NISTIR 7859

ELFT-EFS Evaluation of Latent Fingerprint Technologies:
Extended Feature Sets [Evaluation #2]

M. Indovina
V. Dvornychenko
R. A. Hicklin
G. I. Kiebuzinski

http://dx.doi.org/10.6028/NIST.IR.7859

National Institute of
Standards and Technology
U.S. Department of Commerce

(This page intentionally left blank.)

NISTIR 7859

ELFT-EFS Evaluation of Latent Fingerprint Technologies: Extended Feature Sets [Evaluation #2]

M. Indovina
V. Dvornychenko
Image Group / Information Access Division
Information Technology Laboratory

R. A. Hicklin
G. I. Kiebuzinski
Noblis, Inc.

http://dx.doi.org/10.6028/NIST.IR.7859

May 2012

U.S. Department of Commerce
John Bryson, Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Director

Abstract

The National Institute of Standards and Technology (NIST) Evaluation of Latent Fingerprint Technologies - Extended Feature Sets (ELFT-EFS) consists of multiple ongoing latent algorithm evaluations. This report describes the results and conclusions of ELFT-EFS Evaluation #2; an accuracy test of latent fingerprint searches using features marked by experienced human latent fingerprint examiners, in addition to automatic feature extraction and matching (AFEM). ELFT-EFS Evaluation #1 was the first evaluation of latent fingerprint matchers in which systems from different vendors used a common, standardized feature set. ELFT-EFS Evaluation #2 repeats the same tests using updated matchers. The results show that in most cases there were measureable improvements in the ability of matchers to use images plus manually marked Extended Features (EFS) as an effective and interoperable feature set over the Evaluation #1 results.

The accuracy when searching with EFS features is promising considering the results are derived from early-development matchers. Further studies using next-generation matchers are warranted to determine the performance gains possible with EFS.

Acknowledgements

The authors would like to thank the Department of Homeland Security's Science and Technology Directorate and the Federal Bureau of Investigation's Criminal Justice Information Services Division and Biometric Center of Excellence for sponsoring this work.

Disclaimer

In no case does identification of any commercial product, trade name, or vendor, used in order to perform the evaluations described in this document, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

Technology Providers

This table lists the technology providers who participated in this study. The letter keys listed down the first column are used throughout the report to identify results from specific algorithms. The authors wish to thank the technology providers for their voluntary participation and contribution.

Table ES-1: SDK letter keys and the corresponding technology provider

Key	Technology Provider Name
A	Sagem Securite
B	NEC Corporation
C	3M Cogent, Inc.
D	Sonda Technologies, Ltd.
E	Warwick Warp, Ltd.

Executive Summary

Introduction

The National Institute of Standards and Technology (NIST) Evaluation of Latent Fingerprint Technologies - Extended Feature Sets (ELFT-EFS) consists of multiple ongoing latent algorithm evaluations. This report is an update of ELFT-EFS Evaluation #1 [1] using updated matching algorithms and including minor adjustments to the test database. The test methodology for Evaluation #2 is the same as the one used in Evaluation #1.

The purpose of this test is to evaluate the current state of the art in latent feature-based matching, by comparing the accuracy of searches using images alone with searches using different sets of features marked by experienced latent print examiners. Evaluation #2 provided the vendors the opportunity to correct possible errors and to make adjustments to their algorithms in light of the findings published in Evaluation #1. The feature sets include different subsets of the Extended Feature Set (EFS), which is defined in ANSI/NIST-ITL 1-2011 "Data Format for the Interchange of Fingerprint, Facial, & Other Biometric Information." [2] EFS is a superset of the latent fingerprint feature set currently used by the FBI's Integrated Automated Fingerprint Identification System (IAFIS); EFS features will be used for latent finger and palmprint transactions in the FBI's Next Generation Identification (NGI) system, which is in the process of replacing IAFIS. EFS was developed to serve as an interoperable interchange format for automated fingerprint or palmprint systems, as well as a means of data interchange among latent print examiners. One of the purposes of ELFT-EFS is to determine the extent to which human feature markup is effective. Because human markup is expensive in terms of time, effort, and expertise, there is a need to know when image-only searching is adequate, and when the additional effort of marking minutiae and other features is appropriate. ELFT-EFS is not a test of automatic EFS extraction (i.e. conformance to the standard), but rather a test of data interoperability and how potentially useful such human marked features are when processed by an automated matcher.

The ELFT-EFS evaluations are open to both the commercial and academic community. In Evaluation #2, participants included five commercial vendors of Automated Fingerprint Identification Systems (AFIS). The five participants each submitted three Software Development Kits (SDKs) which respectively contained (i) a latent fingerprint feature extraction algorithm; (ii) algorithms for ten-print feature extraction and gallery creation, and (iii) a 1-to-many match algorithm that returns a candidate list report. The fingerprint features automatically extracted by (i) and (ii) were proprietary, at the discretion of the technology provider, and could include the original input image(s). Evaluations were run at NIST on commodity NIST hardware.

The results from Evaluation #1 were published as NISTIR 7775 [1] and are available at biometrics.nist.gov/cs_links/latent/elft-efs/NISTIR_7775.pdf

Fingerprint and Feature Data

The Evaluation #2 test dataset contained 1,066 latent fingerprint images from 826 subjects. The gallery was comprised of (mated) exemplar sets from all 826 latent subjects, as well as (non-mated) exemplar sets from 99,163 other subjects chosen at random from an FBI provided and de-identified dataset. Each subject in the gallery had two associated exemplar sets: one set of ten rolled impression fingerprint images, and one set of ten plain impression fingerprint images (plains).

In addition to fingerprint images, each latent had an associated set of hand-marked features. The features were marked by twenty-one International Association for Identification Certified Latent Print Examiners (IAI CLPE) using guidelines developed specifically for this process [3]. No vendor-specific rules for feature encoding were used; all encoding was made in compliance with the EFS specification. The various subsets of latent features are summarized in Table ES-2. The additional extended features in subsets LE and LF included ridge quality maps, creases, dots, incipient ridges, ridge edge protrusions, and pores — in addition to the minutiae, ridge counts, cores & deltas, and pattern class included in other subsets. Latent examiners made determinations of Value, Limited Value (latents of value for exclusion only), or No Value at the time of markup, in addition to subjective quality assessments of "Excellent", "Good", "Bad", "Ugly", and "Unusable". A subset of the latents had skeletons marked (including associated ridge flow maps). Features were marked in latent images without reference to exemplars, with the exception of a subset of 418 latent images that included an additional Ground Truth (GT) markup based on the latent and all available exemplars; GT markup provides a measure of ideal (but operationally infeasible) performance when compared to the original examiner markup.

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*3M Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 7**

Table ES-2: Latent feature subsets

Subset	Image	ROI	Pattern class	Quality map	Minutiae + ridge counts	Additional extended features	Skeleton
LA	■						
LB	■	■					
LC	■	■	■	■			
LD	■	■	■		■		
LE	■	■	■	■	■	■	
LF	■	■	■	■	■	■	■
LG		■	■		■		

Primary Findings

The results show that searches using images plus manually marked Extended Features (EFS) demonstrated effectiveness as an interoperable feature set. The four most accurate matchers demonstrated benefit from manually marked features when provided along with the latent image. The latent image itself was shown to be the single most effective search component for improving accuracy, and was superior to features alone in all cases. For all but one matcher, the addition of new EFS features provided an improvement in accuracy. Discrepancies and counter-intuitive behavior noted in Evaluation #1 are no longer evident indicating that possible software errors have been corrected and/or the matcher parameters were adjusted to improve performance. The accuracy when searching with EFS features is promising considering the results are derived from early-development, first-generation implementation of the new standard. Further studies using next-generation matchers are warranted (and underway) to determine the performance gains possible with EFS. In the future further data-specific analyses will be conducted to determine the cases in which specific EFS features provide improvements in accuracy.

Results and Conclusions

1. The highest accuracy for all participants was observed for searches that included examiner-marked features in addition to the latent images.
2. Image-only searches were more accurate than feature-only searches for all matchers.
3. The top performing matchers showed a strong ability to filter out a substantial proportion of false candidates by match score, trading off a moderate drop in accuracy for a very substantial reduction in examiner effort. Since score-based results are more scalable than rank-based results, they provide a better indication of how accuracy would be affected by an increase in database size. This capability could provide important operational benefits such as reduced or variable size candidate lists and greater accuracy for reverse latent searches (searches of databases containing unsolved latents) where a score threshold is used to limit candidate list size.
4. The effect of the use of EFS features other than minutiae is shown in Table ES-3. Bolded results indicate best performance for each matcher. In almost all cases, additional features resulted in accuracy improvement, highlighted in green; cases in which additional features resulted in a decline in accuracy, highlighted in yellow, may be indicative of implementation issues.

Table ES-3: Rank-1 identification rates for latent feature subsets
(Baseline-QA dataset, 418 latents — 100,000 rolled+plain 10-finger exemplar sets)

		Latent Feature Subset						
		LA	**LB**	**LC**	**LD**	**LE**	**LF**	**LG**
Matcher	**A**	63.4	64.1	64.1	**65.6**	**65.6**	64.8	40.4
	B	57.7	60.1	60.1	67.0	67.0	**68.2**	47.4
	C	59.6	60.1	58.6	66.3	**67.2**	*n/a*	45.9
	D	31.8	23.9	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
	E	44.0	46.9	47.1	46.9	47.1	**48.3**	0.0

5. The ground truth (GT) markup method, in which all exemplar mate images were consulted when marking latent features, yielded an increase in performance over the original examiner markup of about 4 to 6 percentage points for image + full EFS searches, and about 12 to 15 percentage points for minutiae-only searches. Though this method is obviously not practical operationally, it shows that matcher accuracy is highly affected by the precision of latent examiner markup, especially in the absence of image data.
6. Latent orientation (angle) has an impact on matcher accuracy. When the orientation of latents was unable to be determined by an examiner, the rank-1 identification rates were substantially (on the order of 20 percentage points) lower than for the latents for which orientation could be determined.
7. Matcher accuracy is very clearly related to the examiners' latent print value determinations, with much greater accuracy for latents determined a priori to be of Value. The matching algorithms demonstrated an unexpected ability to identify low feature content latents: Matcher A's rank-1 accuracy for No Value latents was 20% on image-only searches, and 34.5% on Limited Value latents.
8. The performance of all matchers decreased consistently as lower quality latents were searched, with respect to the subjective scale of "Excellent", "Good", "Bad", "Ugly", or "Unusable".
9. Analysis showed that the greatest percentage of the misses were for latents with low minutiae count, and those assessed by examiners as poor quality ("Ugly"), "No Value" or "Unusable." Algorithm accuracy for all participants was highly correlated to the number of minutiae.
10. At rank 1, 17.8% of the latents in the test were missed by all matchers. Nearly half of these could be individualized by a certified latent examiner.
11. The initial or reviewing examiners determined that 17.6% of the latents in the test could not be used for individualization (Unusable, No Value, of value for exclusion only, or resulted in an inconclusive determination). Nearly half of these were matched by one or more matchers at rank 1.
12. The highest measured accuracy achieved by any individual matcher at rank 1 on any latent feature subset (excluding the use of ground truth markup) was 71.4%, even though approximately 82% of the latents in the test were matched by one or more matchers at rank 1. This indicates a potential for additional accuracy improvement through improved algorithms. The differences in which latents were identified by the various matchers also points to a potential accuracy improvement by using algorithm fusion.
13. All matchers lost or gained a small number of hits as a function of the feature subset used. For high priority cases, where maximum accuracy is desired, it may be worthwhile to submit the search as separate searches using different levels of EFS, (e.g. search using image-only and again search using image+features, and fuse the results).
14. All matchers were more accurate using galleries of both rolled and plain impressions compared to galleries of either rolled or plain impressions separately. And the use of plain impressions in the gallery compared to rolled impressions resulted in a drop in accuracy. For example, searches that included examiner-marked features in addition to the latent images were approximately 6 percentage points more accurate using combined rolled and plain impressions than rolled impressions alone. Similar searches of plain impressions alone were (excluding one outlier) approximately 8 percentage points less accurate than searches of rolled impressions alone.
15. The proportion of the total identifications made by a given matcher that were recorded at rank 1 (IR rank-1 / IR rank-100) is an indication of scalability of performance because identifications at higher ranks are less

likely as gallery size increases. For matchers A, B and C, 87-90% of total identifications are recorded at rank 1 for subset LA; 89-92% of total identifications are recorded at rank 1 for subset LE; 74-79% of total identifications are recorded at rank 1 for subset LG. These results indicate a potential for developing viable candidate list reduction techniques.

16. The "AFIS" markup approach, in which debatable minutiae were removed, was counterproductive in all cases. This result indicates that the matchers are relatively robust when processing debatable minutiae.

Caveats

The performance impact of any specific feature, as measured in this test, may be limited for several reasons: because the participants may not have yet developed approaches to take advantage of such features; because the participants may already take advantage of the underlying information through automated processing of the latent image, and there may be limited or no additional gain from explicitly marking the features; because there was limited opportunity for improvement due to the limited presence of such features in the data; or because explicit human markup of such features may be ineffective for automated matching.

The results may not be applicable to other datasets and operational systems with different processing constraints. Specifically, the relative performance of image-based and feature-based matching may be affected by differences in systems resources. The cost in computer resources may be a significant factor in determining the practicality of image-only or image+feature searches. ELFT-EFS did not measure or evaluate resource requirements for the different types of matching. It should be noted that, as the cost of hardware continues to fall in accordance with Moore's Law, greater use of image-only searches is likely to be more practical.

Table of Contents

Terms and Definitions

This table provides ELFT-specific definitions to various words and acronyms found in this report.

Table 1: Terminology and acronyms

Term	Definition
AFEM	Automated Feature Extraction and Matching
AFIS	Automated Fingerprint Identification System: generic term for large-scale fingerprint matching system
ANSI/NIST	The ANSI/NIST-ITL biometric data standard, used here as the file format containing fingerprint images and features. The current version is ANSI/NIST-ITL 1-2011.
API	Application Programming Interface
CDEFFS	Committee to Define an Extended Fingerprint Feature Set
CMC	Cumulative Match Characteristic
DET	Detection Error Tradeoff characteristic
EBTS	Electronic Biometric Transmission Specification is the standard currently used by FBI for IAFIS, and which will be used for NGI. EBTS is an implementation of the ANSI/NIST-ITL standard. EBTS version 9.3 contains the EFS fields specified in ANSI/NIST-ITL 1-2011. Note there was a name change from Electronic Fingerprint Transmission Specification (EFTS) to EBTS in October 2007. [4]
EFS	Extended Feature Set (proposed extension to ANSI/NIST standards by CDEFFS)
Exemplar	Fingerprint image deliberately acquired during an enrollment process; the known mate of a latent fingerprint
FNIR	False Negative Identification Rate (also called miss rate or false non-match rate)
FPIR	False Positive Identification Rate (also called false match rate)
Fusion	A method of combining biometric information to increase accuracy
Gallery	A set of enrolled ten-prints; synonymous with "database of exemplars." An ELFT Gallery is composed of foreground and background ten-prints.
Ground truth	Definitive association of a print and exemplar at the source attribution ***or*** at the feature level: • ground-truth source attribution is the definitive association of a fingerprint with a specific finger from a subject • feature-level ground truth is the feature-by-feature validation that all marked features (e.g. minutiae) definitively correspond to features present in the exemplar(s).
Hit/hit-rate	A "hit" results when the correct mate is placed on the candidate list; the "hit rate" is the fraction of times a hit occurs, assuming a mate is in the gallery.
IAFIS	The FBI's Integrated Automated Fingerprint Identification System, operational since 1999; for latent prints, IAFIS is scheduled to be replaced by NGI in 2013.
Latent	A fingerprint image inadvertently left on a surface touched by an individual
Matcher	Software functionality which produces one or more plausible candidates matching a search print
Mate	An exemplar fingerprint corresponding to a latent
NGI	The FBI's Next Generation Identification system, which will replace the current IAFIS
NIST	National Institute of Standards and Technology
ROC	Receiver Operator Characteristic
ROI	Region of Interest
Rolled print	A fingerprint image acquired by rolling a finger from side to side on the capture surface
Plain print	A fingerprint image acquired by pressing a finger flat on the capture surface. The plain images used in this test were segmented from slap fingerprints.
Slap print	An image containing multiple plain fingerprints collected simultaneously

1 Introduction

The National Institute of Standards & Technology (NIST) Evaluation of Latent Fingerprint Technology — Extended Feature Sets (ELFT-EFS) is an independently administered technology evaluation of latent fingerprint feature-based matching systems.

The purpose of ELFT_EFS is to evaluate the current state of the art in latent feature-based matching and to assess the accuracy of searches using images alone with searches using different feature sets. A key objective of the evaluations is to determine when human feature markup is effective. Because human markup is expensive in terms of time, effort, and expertise, there is a need to know when image-only searching is adequate, and when the additional effort of marking minutiae and extended features is appropriate.

ELFT-EFS is not a test of automatic EFS extraction (i.e. conformance), but rather a test of data interoperability and how potentially useful such human features are when processed by a matcher. When images are included in a search, automatic feature extraction may be used to any degree participants choose (e.g. in addition to or in place of the manually specified EFS features): which features are automatically extracted from latent or exemplar images are at the sole discretion of the participant and are not examined in this study.

This report covers the results and conclusions of ELFT-EFS Evaluation #2, which follows the prior Evaluation #1 study [1]. As in Evaluation #1, the feature sets used in Evaluation #2 include different subsets of the Extended Feature Set (EFS) features incorporated in ANSI/NIST-ITL 1-2011 [3].

For an exact description of the evaluation data, test protocol, and EFS feature subsets used for ELFT-EFS, refer to NISTIR 7775 [1].

Following completion of the ELFT-EFS Evaluation #1, participants were given an opportunity to view the latents that their matcher was unable to identify at rank-1 for any latent subset. All such latents are considered "misses" for the purposes of failure or "miss" analysis. For all such "miss" cases, participants were permitted to (i) view the latent image; (ii) view all associated feature markup; (iii) view the rolled and plain exemplar mates; and (iv) view, when available, the rank and score of the exemplar mate. All images were viewed on NIST hardware by the participants at NIST, and direct access to the images was not provided. One exception to this were the latent and exemplar images from the MLDS dataset (see Section 3.1.1 of NISTIR 7775 [1]) which were provided to the participants along with associated feature markup, and rank and score results. For this reason, the MLDS latents were excluded from the scoring of the Evaluation #2 results, though they are available in Appendix A.

Participants in Evaluation #2 were given the choice of which subsets to be tested on. Most chose to be re-tested on all subsets used in Evaluation #1, with the following exceptions: 3M Cogent chose not to be re-tested on subset LF; Sonda chose to be re-tested only on subsets LA-LB.

ELFT Phase II

ELFT Phase II was a test of AFEM-only matching conducted using participants' software on NIST hardware at NIST facilities. The vast majority of the data used in ELFT Phase II was sourced from successful operational searches of the FBI's IAFIS. ELFT Phase II results were published in April 2009 [5].

ELFT-EFS Public Challenge

Prior to ELFT-EFS Evaluation #1, the ELFT-EFS Public Challenge was conducted as a practice evaluation on public data to validate formats and protocols. The results of the Public Challenge are included as Appendix B of ELFT-EFS Evaluation #1 [1].

ELFT-EFS Evaluation #1

The ELFT-EFS Evaluation #1 was conducted using participants' software on NIST hardware at NIST facilities. Datasets were from multiple sequestered sources, each broadly representative of casework. The ELFT-EFS Evaluation #1 was run specifically to identify any near-term benefits, NOT to identify long-term feasibility/accuracy. Timing constraints, subtests, and analysis were based in part on the results and lessons learned from the ELFT-EFS Public Challenge. Participation in the public challenge was a prerequisite for participation in Evaluation #1. ELFT-EFS Evaluation #1 results were published in March 2011 [1].

ELFT-EFS Evaluation #2

The ELFT-EFS Evaluation #2 uses the test data, API, and test protocols used in Evaluation #1. ELFT-EFS Evaluation #2 was conducted with the expectation that participants from Evaluation #1 would incorporate lessons learned based on the Evaluation #1 report and the miss analysis review conducted at NIST.

Subsequent Evaluations

Subsequent ELFT-EFS Evaluations are expected to be conducted to evaluate different aspects of latent matching, respond to lessons learned, and track ongoing progress.

A detailed description of the evaluation may be found in the Evaluation #1 Test Plan, which is in Appendix A of NISTIR 7775 [1].

2 Participants

The ELFT-EFS evaluations are open to both the commercial and academic community. In Evaluation #2, the participants included all five vendors of Automated Fingerprint Identification Systems (AFIS) who participated in Evaluation #1: Sagem Securite, NEC, 3M Cogent, Sonda, and Warwick (see Table ES-1). An additional vendor, SPEX Forensics, has also participated in Evaluation #2 however their results are not available at the time of this publication and will be reported in a follow-on publication.

3 Methods of analysis

Analyses of the accuracy of 1:N identification searches returning candidate lists can be with respect to rank or score.

3.1 Rank-based analyses

Identification rate at rank *k* is the proportion of the latent images whose mate is reported at rank *k* or lower in the search candidate list. Identification rank ranges from 1 to 100, as 100 was the (maximum) candidate list size specified in the API.

Overall accuracy results for rank-based metrics are presented via Cumulative Match Characteristic (CMC) curves. A CMC curve shows how many latent images are correctly identified at rank 1, rank 2, etc. A CMC is a plot of identification rate (also known as "hit rate") vs. recognition rank.

Rank-based analyses are specific to the gallery size used in the test, and cannot be assumed to scale to substantially larger gallery sizes.

3.2 Score-based analyses

The True Positive Identification Rate (TPIR) indicates the fraction of searches where an enrolled mate exists in the gallery in which enrolled mates appear in the top candidate list position (i.e. rank 1) with a score greater than the a given threshold. (Note that the False Negative Identification Rate (FNIR = 1-TPIR) indicates the fraction of searches in which enrolled mates do not appear in the top position with a score greater than the threshold.)

The False Positive Identification Rate (FPIR) indicates the fraction of candidate lists (without enrolled mates) that contain a non-mate entry in the top candidate list position with a score greater than a given threshold.

In theory, analysis could use a combination of score and rank, in which scores are filtered based on rank. In practice, score-based results at rank 1 and at rank 100 were not notably different, so results presented are for scores at rank 1.

Score-based results are of interest for multiple reasons:

- Score-based results are more scalable than rank-based results, providing a better indication of how accuracy would be affected by an increase in database size. As a rule of thumb, the TPIR at FPIR = 0.01 provides a rough projection of accuracy for an increase in database size of 100x.
- For reverse or unsolved latent matching, in which a gallery of latents is searched with newly acquired exemplars, potential candidates must be automatically screened to limit the impact on human examiners. Score-based results give an indication of the potential effectiveness of reverse matching.

For score-based results, derivations of Receiver Operating Characteristic (ROC) curves were plotted using the methodology defined in ELFT Phase II ([5], Section 3.1.2, p 24.) [a]. All ROC curves in this analysis are limited to Rank 1 (limited to the highest scoring result in the candidate list), based on findings from Evaluation #1 that ROC results did not show any notable effect when limited to rank 1 as opposed to higher ranks. A horizontal line in an ROC indicates no degradation in accuracy when non-mates are excluded.

Note that when FPIR=1.0, the score-based TPIR is the same as the rank-1 identification rate shown in the rank-based (CMC) analyses shown in section 5.

In each case, participants returned a raw score and a normalized score estimating the probability (1-100) of a match. The normalized scores provided equal or better results than the raw scores and therefore only the normalized results are reported here.

4 Data

The Evaluation #2 test dataset contained 1,066 latent fingerprint images from 826 subjects. The gallery was comprised of (mated) exemplar sets from all 826 latent subjects, as well as (non-mated) exemplar sets from 99,163 other subjects chosen at random from an FBI provided and de-identified dataset. Each subject in the gallery had two associated exemplar sets: one set of ten rolled impression fingerprint images, and one set of ten plain impression fingerprint images (plains). The Evaluation #1 test dataset contained 1,114 latent fingerprint images, of which forty-eight (48) were removed for Evaluation #2: 38 latents (from 4 subjects) were removed from the dataset since they were provided as example images to the participants for miss-analysis purposes following Evaluation #1; 10 of the Evaluation #1 latents that did not have mates in the gallery were also removed.

4.1 Latent Data

4.1.1 Sources of latent images

The latent images came from both operational and laboratory collected sources, as shown in Table 2. Each of the initial data sources included a larger number of latent images. From these sources, examiners provided assessments of the quality of the latents using a subjective "Excellent", "Good", "Bad", "Ugly" and "Unusable" scale. The latents selected for subsequent markup included approximately equal proportions of the Good, Bad, and Ugly categories, with less than 2% of each of the Excellent and Unusable categories. (See Section 4.1.3) In none of the cases were the mates selected through the use of automated fingerprint matchers (possibly producing AFIS bias), as was true in the ELFT Phase 2 evaluation. Each latent image was from a distinct finger.

Table 2: Sources of latent images *(Baseline dataset)*

Name	# Latents	Description	Minutiae count	
			Mean	St dev
Casework 1	368	Operational casework images	20	12
Casework 2	165	Operational casework images	18	9
WVU	440	Laboratory collected images	27	19
FLDS	93	Laboratory collected images	20	17
Total (Baseline)	1066		22	16

A combination of prints from both operational casework and laboratory collected images was included to provide a diversity of data types, including a broad range of quality, deposition, and substrate/background attributes. All of the prints in the casework datasets were considered of value by the original examiners who determined the individualizations.

[a] *Note that ROC curves are used here instead of the DET curves used in ELFT Phase II. ROCs and DETs display the same information with the sole difference that ROCs display the true positive rate on the Y axis, while DETs display the inverse (the false negative/type-2 error rate is 1-true positive rate) on the Y axis, generally in log scale. DETs are effective at showing distinctions between small error rates, but are more difficult to interpret than ROCs for the accuracy levels reported here.*

For an evaluation such as this, source attribution, or "ground truth"[b] association of the latent with a specific finger of a known subject, is critical. The source attribution of the datasets was accomplished differently for the casework and laboratory-collected sources. The laboratory-collected images were acquired under carefully controlled conditions so that the source attribution could not be in doubt. In the casework datasets, the prints used were selected from cases in which multiple additional corroborating latents and exemplars were used in the individualization, so that the latent-exemplar relation was not made solely through the use of these latents.

The technology providers had no knowledge of, or access to, the test datasets prior to or during the tests, other than a small set of public fingerprints provided to serve as examples – the Multi-Latent Dataset (MLDS) discussed in Section 1 which was excluded from the results of Evaluation #2.

4.1.2 Latent features

In addition to fingerprint images, each latent had an associated set of hand-marked features. The features were marked by twenty-one International Association for Identification Certified Latent Print Examiners (IAI CLPE). Latent features were included in ANSI/NIST files formatted in accordance with the Extended Feature Set fields defined in ANSI/NIST ITL-1 2011 "Data Format for the Interchange of Fingerprint, Facial, & Other Biometric Information." [3] EFS is a superset of the latent fingerprint feature set used by the FBI's Integrated Automated Fingerprint Identification System (IAFIS). In marking the latent features, examiners followed the specific guidelines defined in [3]; no vendor specific rules for feature encoding were used. The test evaluated different combinations of EFS fields. The different subsets of EFS features included in the latent files (Subsets LA-LG) are defined in Table 3. The specific EFS fields included in each subset are listed in Appendix A of NISTIR 7775 [1]. The additional EFS features in subsets LE and LF included ridge quality maps, creases, dots, incipient ridges, ridge edge protrusions, and pores — in addition to minutiae, ridge counts, cores & deltas, and pattern class. Latent examiners made determinations of Value, Limited Value (latents of value for exclusion only), or No Value at the time of markup, in addition to subjective quality assessments of "Excellent", "Good", "Bad", "Ugly", and "Unusable".

Table 3: Latent feature subsets

Subset	Image	ROI	Pattern class	Quality map	Minutiae + ridge counts	Additional extended features	Skeleton
LA	■						
LB	■	■					
LC	■	■	■	■			
LD	■	■	■		■		
LE	■	■	■	■	■	■	
LF	■	■	■	■	■	■	■
LG		■	■		■		

Features were marked in latent images without reference to exemplars, with the sole exception of a subset of the latent images that included an additional Ground Truth (GT) markup based on the latent and all available exemplars discussed in section 8. GT markup provides a measure of ideal (but operationally infeasible) performance when compared to the original examiner markup. GT data eliminates the variability in feature identification introduced by the latent examiner. GT results represent the upper performance limit of what the matching algorithm can achieve.

Note that conformance testing of automatic extraction of EFS features was not part of this test. In other words, the evaluation did not measure how close automatically extracted features were to examiner created features. The extent of use of the EFS features for the test was solely decided by the participants. However, the EFS feature set was presented to all vendors to use as they saw fit. Automated algorithms can use the extended features defined for a latent search without explicitly computing them for the exemplar image, and thus it must be emphasized that automated extraction of the extended features on the exemplar is not necessarily the only, nor the best way, to use this information. For example, an examiner may mark an area as a scar; for the exemplar, the matcher would not necessarily have to mark the area as a scar, but may use that image based information to match against a corresponding area that would otherwise have many "false" minutiae and poor ridge flow.

[b] *Note that the term "ground truth" is commonly used in two contexts: source attribution (the definitive association of a fingerprint with a specific finger from a subject), or feature-level ground truth (the feature-by-feature validation that all marked features (e.g. minutiae) definitively correspond to features present in the exemplar(s)).*

All latent markup was not available for all of the latents: skeletons were only available for a subset due to the extensive time required for manual markup of skeletons. The full 1,066 dataset is named here as the "Baseline" dataset. The 418 latent subset of Baseline that includes skeletons is named here as the "Baseline-QA" dataset. Because skeletons were only available for Baseline-QA, the Baseline-QA dataset was used to compare all of the latent feature subsets (LA-LG). Baseline results were only used to compare LA, LE, and LG.

In addition to the standard markup approach described throughout this document, the 418 latents in Baseline-QA were marked up using two additional markup methods to assess the effects of a more conservative markup approach ("AFIS") and ideal (but operationally infeasible) Ground Truth markup ("GT") – see Table 15.

A quality assurance review was performed by additional latent examiners to verify that the markup for each latent was acceptable. When the reviewing examiner did not believe that the markup was accurate (due to factors such as missing minutiae or incorrect pattern class), the marked up latent was returned to the original examiner with instructions to review and correct feature markup, without specifically referring to individual features of concern; when the original examiner was not available, latents with known markup issues were removed from the dataset.[c]

Examiners were instructed to mark all features present in each latent (with the exception of skeletons, which were only marked in a specified subset of the latents). Each latent had a non-zero number of minutiae, and ridge quality maps were marked in all latents. The prevalence of the other features is summarized in Table 4, which shows the proportion of the latents in the Baseline dataset with any of the specified features marked; distributions are detailed in Appendix C of NISTIR 7775 [1].

Table 4: Presence of additional extended features

Feature type	% of Baseline with features present	Feature count (when present)	
		Mean	St dev
Dots	21%	3.0	4.9
Ridge edge features	4%	3.7	7.1
Distinctive features (e.g. scars)	3%	1.1	0.4
Incipient ridges	19%	5.1	9.2
Pores	71%	138.9	176.7
Local quality issues	61%	1.9	1.3
Creases	31%	4.7	5.7
Cores	75%	1.3	0.5
Deltas	41%	1.1	0.3

4.1.3 Value and quality assessments of latents

When latent images were selected for markup, the latents were assessed by an examiner using an subjective quality scale of "Excellent", "Good", "Bad", "Ugly", and "Unusable". Subsequently, the (different) examiners who marked the data made value determinations at the time of markup, using the categories defined in EFS [Field 9.353, Examiner value assessment]:

- Value: The impression is of value and is appropriate for further analysis and potential comparison. Sufficient details exist to render an individualization and/or exclusion decision.
- Limited: The impression is of limited, marginal, value and may be appropriate for exclusion only.
- No value: The impression is of no value, is not appropriate for further analysis, and has no use for potential comparison.

Table 5 shows the counts and proportion of each of the value and quality determinations in the Baseline dataset. Note that 2.4% of the latents in Baseline were marked as No Value, and an additional 10.6% were marked as Limited value. The prints of Limited or No Value were included so that the capabilities of matchers could be tested on very poor-quality prints; all prints included at least one marked minutia. Since different examiners made the quality and value assessments, there is variation, most notably in the No Value and Unusable categories, and in the fact that some latents were assessed as both Good and Limited or No Value.

[c] *To be explicit: the only data removed from the dataset was done on the basis of inaccurate markup by examiners, not due to any attributes of the images.*

Table 5: Comparison of quality and value determinations ***(Baseline dataset)***

		Value assessment					
		Of Value	Limited Value	No Value	*n/a*	Total	%
Quality assessment	**Excellent**	8				8	0.8%
	Good	312	4	1	*6*	323	30.3%
	Bad	330	16	2	*3*	351	32.9%
	Ugly	243	88	6	*2*	339	31.8%
	Unusable	1	3	14		18	1.7%
	n/a	*23*	*2*	*2*		*27*	*2.5%*
	Total	917	113	25	*11*	1066	
	%	86.0%	10.6%	2.4%	*1.0%*		

Table 6 shows the relationship between quality/value determination and minutiae count. The minutiae quality was determined by the EFS ridge quality map field, which differentiates between areas with definitive and debatable minutiae. As expected, both quality and value determination are highly correlated with minutiae count, as well as with the count and proportion of definitive minutiae. The relationships between minutiae count, value determination, and quality with respect to accuracy are reported in Sections 11-13.

Table 6: Minutiae count statistics by quality and value determination ***(Baseline dataset)***

	Definitive minutiae (mean)	Debatable minutiae (mean)	Definitive minutiae as % of total minutiae
All (Baseline)	13.5	8.9	60%
Excellent	54.1	14.0	79%
Good	24.6	9.3	73%
Bad	11.4	9.8	54%
Ugly	5.3	7.7	41%
Unusable	0.0	3.9	0%
Of Value	15.3	9.5	62%
Limited Value	2.2	5.5	29%
No Value	0.3	3.3	9%

4.1.4 Orientation

Latent fingerprint images varied in orientation from upright ±180°. Table 7 shows the distribution of the Baseline latents by orientation, as determined by latent examiners during markup. The relationship between latent orientation and accuracy is reported in Section 10.

Table 7: Orientation of latents ***(Baseline dataset)***

Orientation (degrees from upright)	% of Baseline latents
Unknown	7.0%
0-9 degrees	56.4%
10-19 degrees	12.1%
20-44 degrees	18.4%
45-89 degrees	5.1%
>90 degrees	1.0%

4.2 Galleries

The galleries against which the latent datasets were searched included the combinations of rolled and plain impressions specified in Table 8. Unless otherwise noted, all results in this report are against the rolled + plain (E1) subset of exemplars. Exemplars came from optical livescan and inked paper sources. An exemplar record for one

subject always included all ten fingers. The relationship between exemplar impression type and accuracy is reported in Section 7.

Table 8: Gallery subsets

Exemplar subset	# subjects	Description
rolled + plain (E1)	100,000	10 rolled & 10 plain impressions each
rolled (E2)	10,000	10 rolled impressions each
plain (E3)	10,000	10 plain impressions each

Plain impressions were segmented from slap images. For the non-mated data, the slap segmentation was performed automatically; for the exemplars mated to the latent probes, human review was conducted to verify the accuracy of segmentation. Exemplar images were retained in the same orientation as they were captured, including the segmented slap images; this was conveyed to the participants in the Test Plan.

Of the non-mated exemplars, approximately 54% were from live-scan sources, and 46% were from ink. Of the mated exemplars, approximately 48% were from live-scan sources, and 52% were from ink.

No feature markup data was provided for the exemplar images.

The number of subjects in the gallery was selected to be as large as possible given finite testing resources and throughput requirements (see Section 4 of NISTIR 7775 [1]).

4.3 Data Format

All images and EFS feature markup data were contained in ANSI/NIST files as described in Appendix A of NISTIR 7775 [1].

All images were 8-bit grayscale. All latent images were 1000 pixels per inch (39.37 pixels per millimeter (ppmm)), uncompressed. Exemplar images were 500 pixels per inch (19.69 pixels per millimeter (ppmm)), compressed using Wavelet Scalar Quantization (WSQ) [6].

Latent fingerprint images varied from 0.3″x 0.3″ to 2.0″ x 2.0″ (width x height).

Exemplars were provided in complete 10-finger sets, with finger positions noted. The finger positions for latents were not noted – no searches were restricted to specific fingers.

5 Rank-based results

The following tables summarize the rank-1 identification rates for each of the matchers for each of the latent subsets as searched against exemplar set E1. Table 9 shows summary results for the Baseline-QA dataset, the (418 latent) subset of the (1,066 latent) Baseline dataset for which skeletons were available; for that reason, the Baseline-QA dataset is used to compare all of the latent subsets. Bolded results indicate best performance for each matcher. Cases in which additional features resulted in accuracy improvement are highlighted in green; cases where accuracy declined are highlighted in yellow.[d] Performance generally increases as more information is added. While subsets LA through LF each contain an increasing amount of information, LG is not part of that sequence, and actually contains the least amount of information. In interpreting these results it should be kept in mind that

[d] *This is based on the following superset relationships (see also Table 3):*

Latent subset	*Is a superset of the features in*
LA	-
LB	*LA*
LC	*LB*
LD	*LB (or LG)*
LE	*LD (or LC)*
LF	*LE*
LG	-

statistical random fluctuations are on the order of 5% [e]. The greatest improvements were achieved by matchers C ,D and E.

Note: in this section and throughout the report, results from Evaluation #1 differ slightly from those reported in NISTIR 7775, because these are based on the revised Baseline and Baseline-QA datasets. In general, the Evaluation #1 results reported here are about 1 percentage point higher than those reported in NISTIR 7775, because 10 (0.9%) of the Evaluation #1 latents that did not have mates in the gallery were removed.

Table 9: Rank-1 identification rates for latent feature subsets
(Baseline-QA dataset, 418 latents — 100,000 rolled+plain 10-finger exemplar sets)

Table 9A – ELFS-EFS Evaluation #1 results

		Latent Feature Subset						
		LA	**LB**	**LC**	**LD**	**LE**	**LF**	**LG**
Matcher	**A**	60.8	59.3	59.3	**63.9**	**63.9**	63.2	41.2
	B	57.9	58.9	59.1	60.5	60.5	**62.0**	45.7
	C	41.2	41.9	44.0	58.9	60.3	**60.8**	44.7
	D	22.5	*n/a*	*n/a*	13.6	*n/a*	**14.4**	10.1
	E	43.8	44.7	47.1	45.2	**48.3**	31.8	23.9

Table 9B – ELFS-EFS Evaluation #2 results

		Latent Feature Subset						
		LA	**LB**	**LC**	**LD**	**LE**	**LF**	**LG**
Matcher	**A**	63.4	64.1	64.1	**65.6**	**65.6**	64.8	40.4
	B	57.7	60.1	60.1	67.0	67.0	**68.2**	47.4
	C	59.6	60.1	58.6	66.3	**67.2**	*n/a*	45.9
	D	31.8	23.9	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
	E	44.0	46.9	47.1	46.9	47.1	**48.3**	0.0

Table 10 shows the rank-1 results for the complete Baseline dataset. Skeletons (LF) were not marked for the complete dataset. Only the subsets LA, LE, and LG were run for the complete Baseline dataset. LB/LC were omitted because they showed limited improvement over LA, and LD was omitted because the performance of LD and LE were so similar. Bolded results indicate best performance for each matcher.

The performance for the complete Baseline dataset is better than Baseline-QA by about 4-5% in most cases. This is an incidental artifact of the process by which Baseline-QA was selected, which resulted in a greater proportion of low-quality latents. Therefore, caution should be used in comparing the results for Baseline and Baseline-QA.

[e] *Approximate, based on the binomial distribution and 95% confidence interval.*

Table 10: Rank-1 identification rates for latent feature subsets
(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

Table 10A – ELFS-EFS Evaluation #1 results

		Latent Feature Subset		
		LA	**LE**	**LG**
Matcher	**A**	63.3	**67.7**	44.7
	B	62.5	**64.5**	49.2
	C	49.2	**63.1**	48.6
	D	25.5	16.7	11.7
	E	48.2	**51.2**	29.6

Table 10B – ELFS-EFS Evaluation #2 results

		Latent Feature Subset		
		LA	**LE**	**LG**
Matcher	**A**	67.2	**70.2**	45.1
	B	63.0	**69.9**	49.8
	C	65.0	**71.4**	49.3
	D	38.9	*n/a*	*n/a*
	E	49.2	**52.3**	0.0

The statistical random fluctuations are on the order of 3% (lower than previous because of a larger sample size).[f] Thus, many of the cases show significant statistical improvement, especially Matcher C. In one case (E on LG) the new version of the matcher does not appear to operate properly. Image-only searches (LA) were far more accurate for matchers A, B and E than minutiae-only searches (LG). This is a notable result because the performance of 1990s AFIS would have resulted in expectations that the reverse would have been true. It should be noted that relative performance of image-based and feature-based matching may be affected by differences in systems resources, and therefore may differ among evaluations.

In general, matchers showed a net improvement when features were added, most notably between LA and LD | LE. It is interesting to note that the average improvement between LA and LE (discounting Matcher D) was 5.8% for Evaluation #1, but only 4.9% for Evaluation #2. However, the main reason for this drop was the huge improvement in image-only (LA) performance achieved by Matcher C.

Overall accuracy is affected by the quality distribution of the latent and exemplar prints (see Sections 12-13); as discussed above, 13% of the Baseline dataset was marked as Limited or No Value. Accuracy will differ given a different distribution of poor-quality latents.

The following graphs show the complete CMCs for the rank-1 through rank-100 results.

Observations (all CMC graphs):

- The CMC curves for the different matchers are generally quite parallel, so that the difference in performance between two matchers or subsets does not differ substantially with respect to rank.
- The CMCs are relatively flat after rank 20: the number of additional candidates in excess of 20 is very small for the top performing matchers.

[f] *Approximate, based on the binomial distribution and 95% confidence interval.*

- The proportion of the total identifications made by matchers that were recorded at rank 1 ($IR_{rank\text{-}1}$ / $IR_{rank\text{-}100}$) is an indication of scalability of performance because identifications at higher ranks are less likely as gallery size increases. For matchers A, B and C, 87-90% of identifications are recorded at rank 1 for subset LA; 89-92% of identifications are recorded at rank 1 for subset LE; 74-79% of identifications are recorded at rank 1 for subset LG. The trend shows that as the information used for the search increases, identifications are made at higher ranks.[8]
- Performance using "features only" (dataset LG) is significantly less than for the other sets. For matchers A, B, C, and E, substantial improvement (20-25%) is shown for LE over the legacy minutiae-only feature set (LG).
- Matcher A performance for LA searches showed the least performance change for the Baseline and Baseline-QA datasets, unlike the other matchers. Since the Baseline-QA dataset had a greater number of poor quality prints, this result may indicate that A is more robust when presented with poor quality image data (see also Section 12).

[8] *The proportion of times in a typical candidate list that the correct mate was recorded at rank 1, as opposed to ranks 2-20 (IR rank-1 / IR rank-20,) for matchers A, B, and C was: 90-92% for LA ; 91-94% for LE; and 81-84% for LG.*

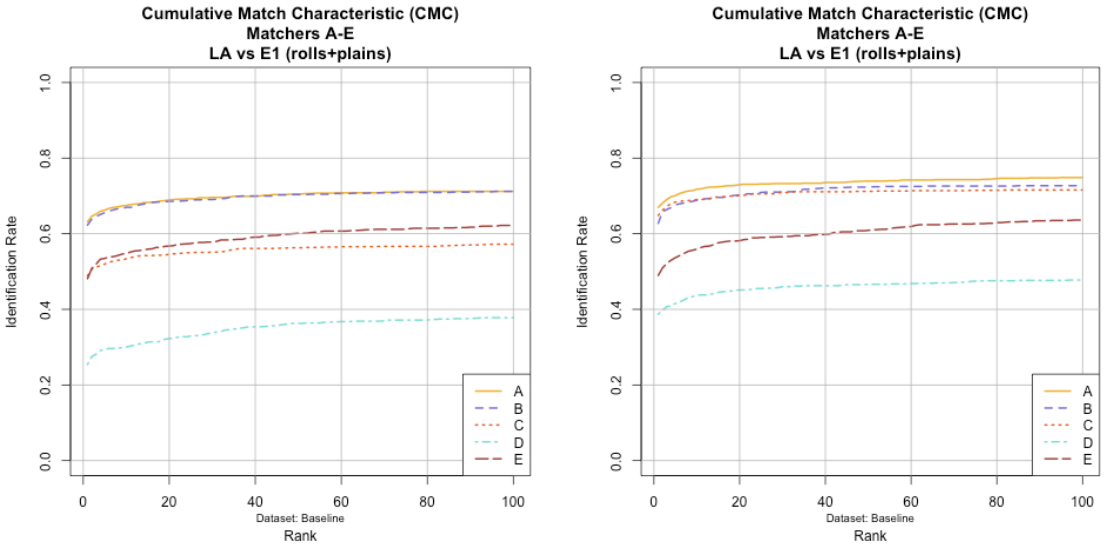

Figure 1A: Performance Comparison, LA vs. E1 – Evaluation #1 on left, #2 on right

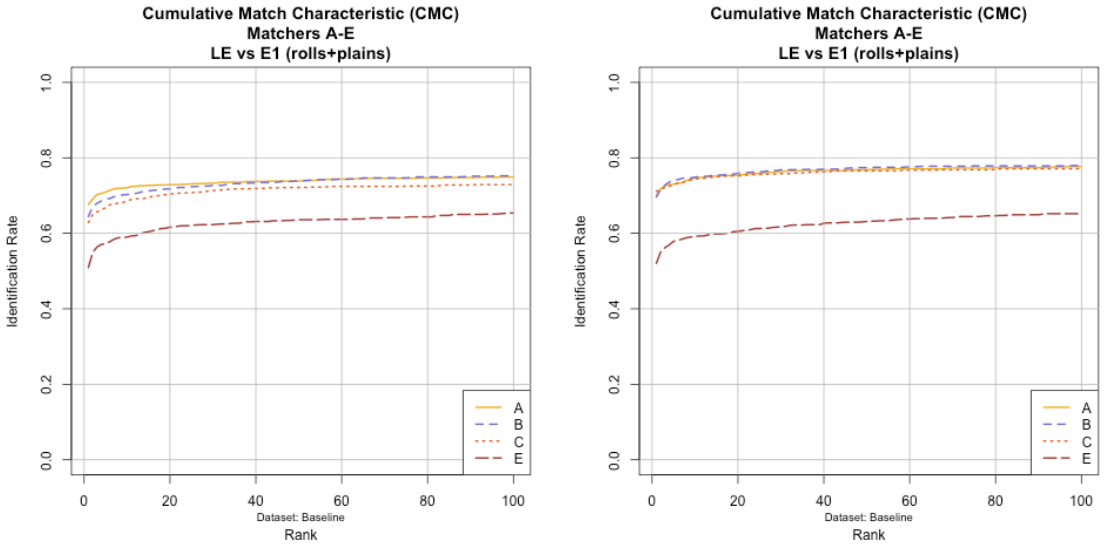

Figure 1B: Performance Comparison, LE vs. E1 – Evaluation #1 on left, #2 on right

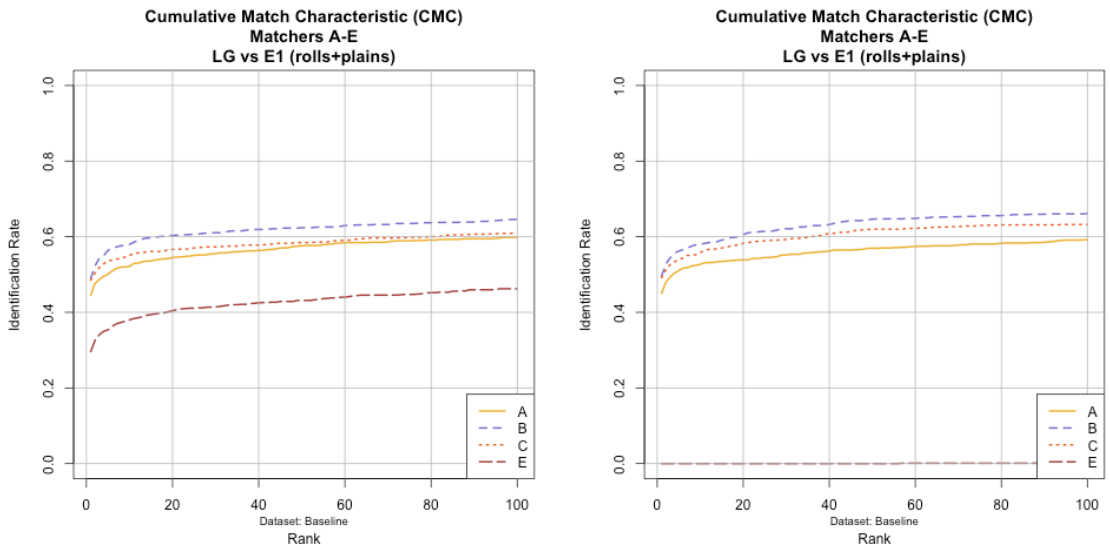

Figure 1C: Performance Comparison, LG vs. E1 – Evaluation #1 on left, #2 on right

Figure 1 (parts A,B,C): Rank-based comparison of matchers for latent subsets LA, LE and LG

(Baseline dataset, 1,066 latents — 100,000 rolled+plain 10-finger exemplar sets)

Figure 2 shows how each of the matcher performed on the individual datasets. These results are based on the Baseline-QA datasets since not all feature subsets were included for the larger Baseline set. As in the previous series, the left graph shows the Evaluation #1 results, while the right graph shows Evaluation #2.

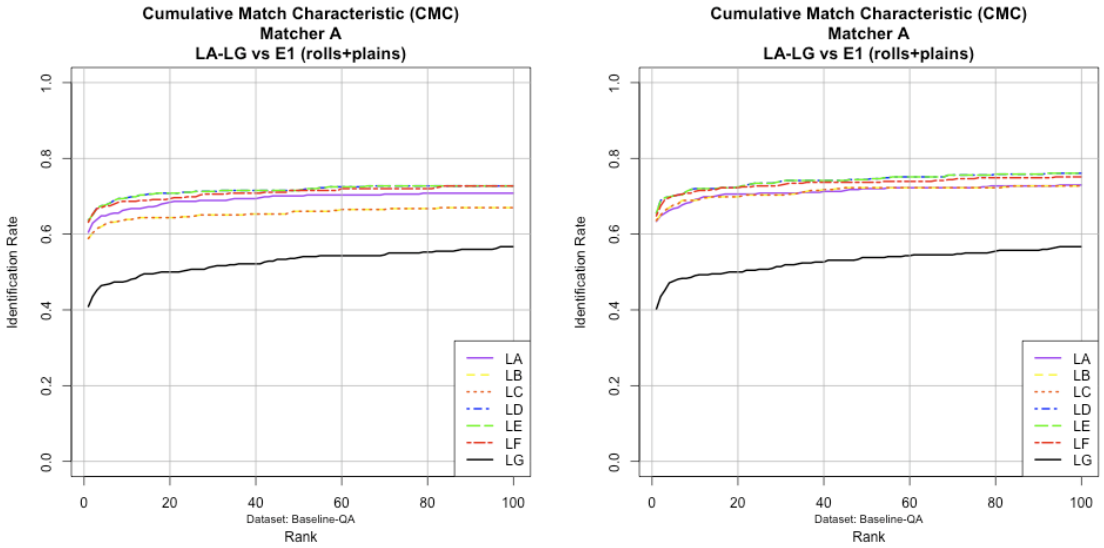

Figure 2A – Performance of Matcher A on LA-LG vs. E1– Evaluation #1 on left, #2 on right

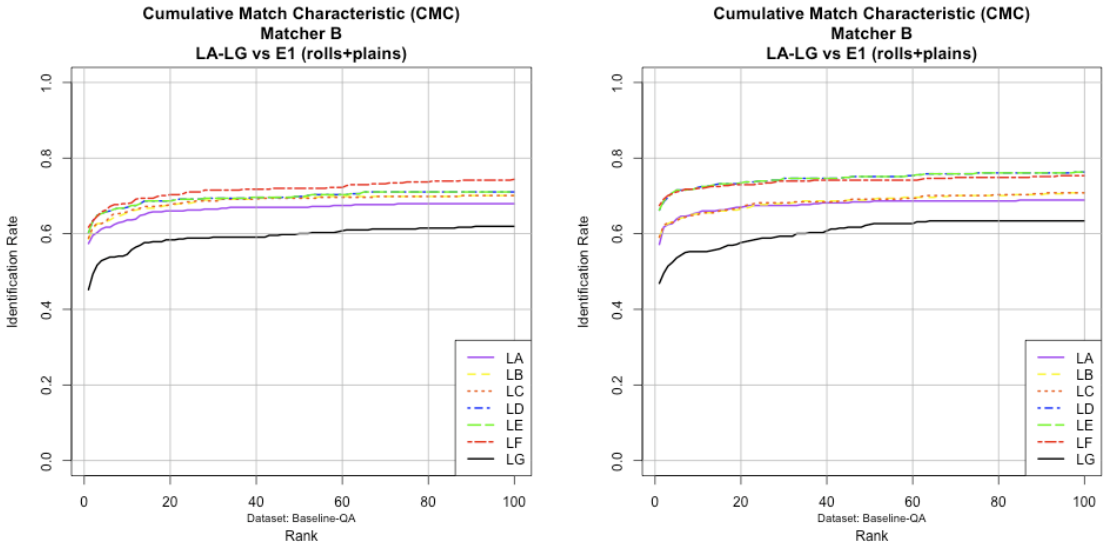

Figure 2B – Performance of Matcher B on LA-LG vs. E1 – Evaluation #1 on left, #2 on right

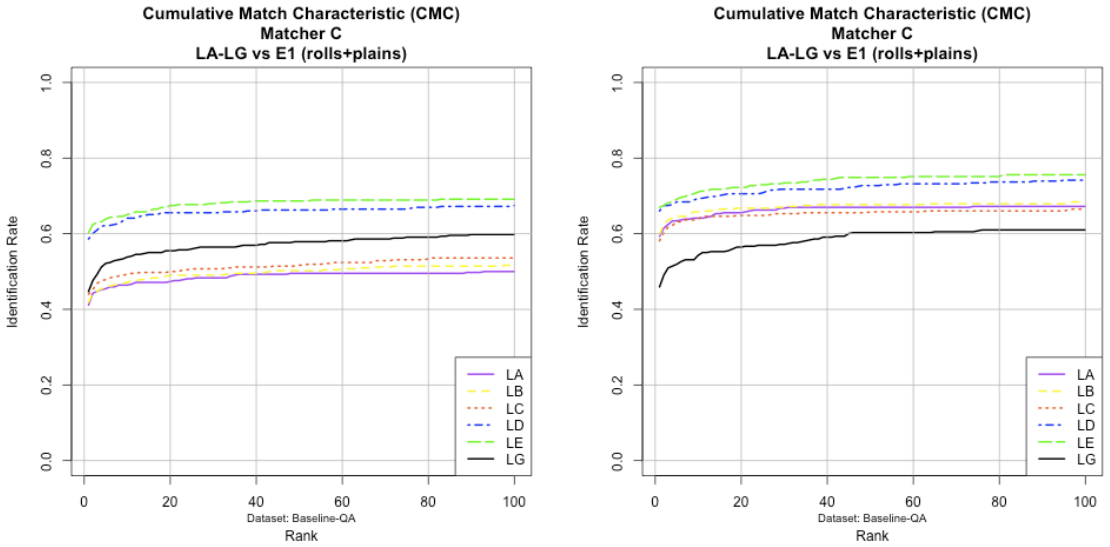

Figure 2C – Performance of Matcher C on LA-LG vs. E1– Evaluation #1 on left, #2 on right

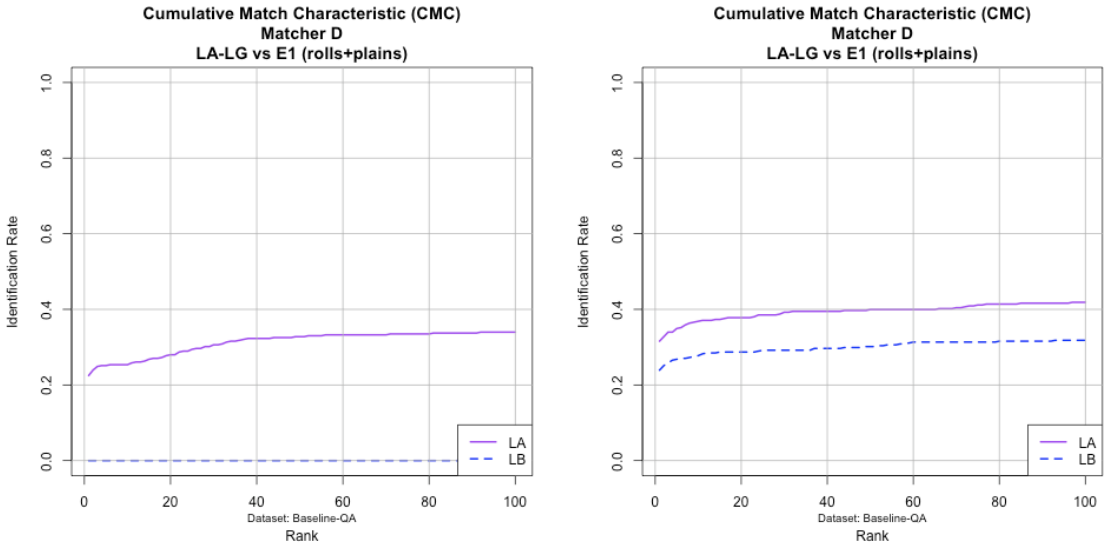

Figure 2D – Performance of Matcher D on LA-LG vs. E1– Evaluation #1 on left, #2 on right

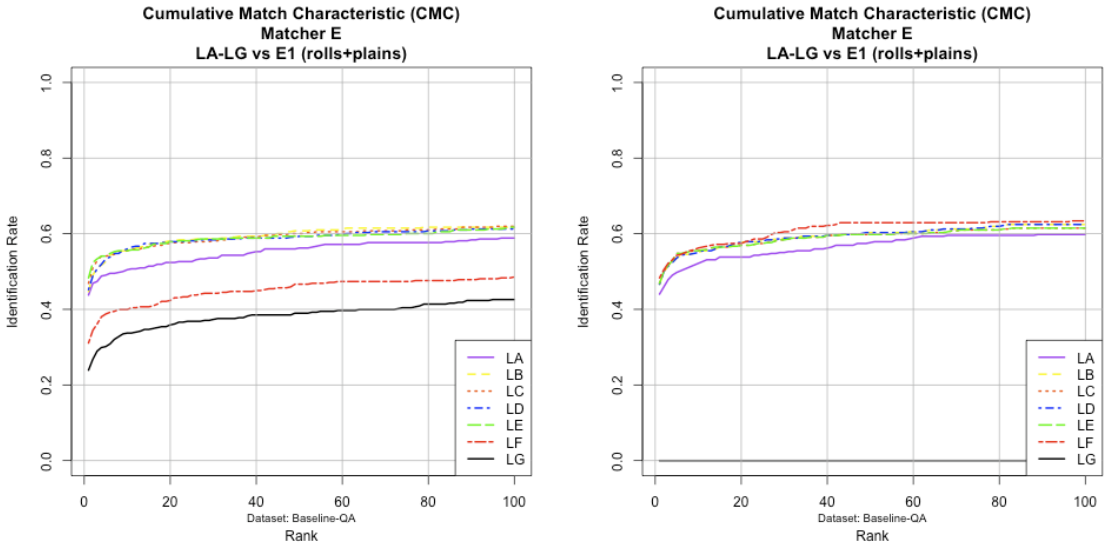

Figure 2E – Performance of Matcher E on LA-LG vs. E1– Evaluation #1 on left, #2 on right

Figure 2 (parts A-E): Rank-based comparison of matchers A-E for latent subsets LA-LG
(Baseline-QA dataset, 418 latents — 100,000 rolled+plain 10-finger exemplar sets)

The majority of cases show improvements between Evaluation #1 and Evaluation #2, particularly for Matchers C and E. In a few isolated cases performance was worse for Evaluation #2; for example, Matcher E failed to operate properly on LG.

6 Score-based results

The ROC charts in Figures 3 and 4 show the effect of automatically filtering candidates based on score, ignoring all candidates above rank 1. In the ROCs below, the rightmost point of each curve is identical to the rank-1 result for the corresponding CMC in Section 5. Figure 3B shows that matcher A has a rank-1 IR of 0.65, and the TPIR remains at 0.65 when moving from FPIR=1.0 to FPIR=0.5: this means that if a score threshold were used to filter results, about half of the candidate lists that did not include a true mate could be eliminated without any impact on

accuracy. If the score threshold is set to eliminate 99% of the candidate lists (FPIR=0.01), the TPIR for matcher A would drop from 0.65 to 0.57, trading off a moderate drop in accuracy for a very substantial reduction in examiner effort.

Note that for some matchers the curves do not extend fully across the charts; this simply means that the matcher scores did not fully populate the range of FPIR. The matchers with the higher TPIR at the lower values of FPIR can also be expected to maintain a higher identification rate as gallery size increases. In general, the flatter the curve, the less sensitivity to increases in the gallery size.

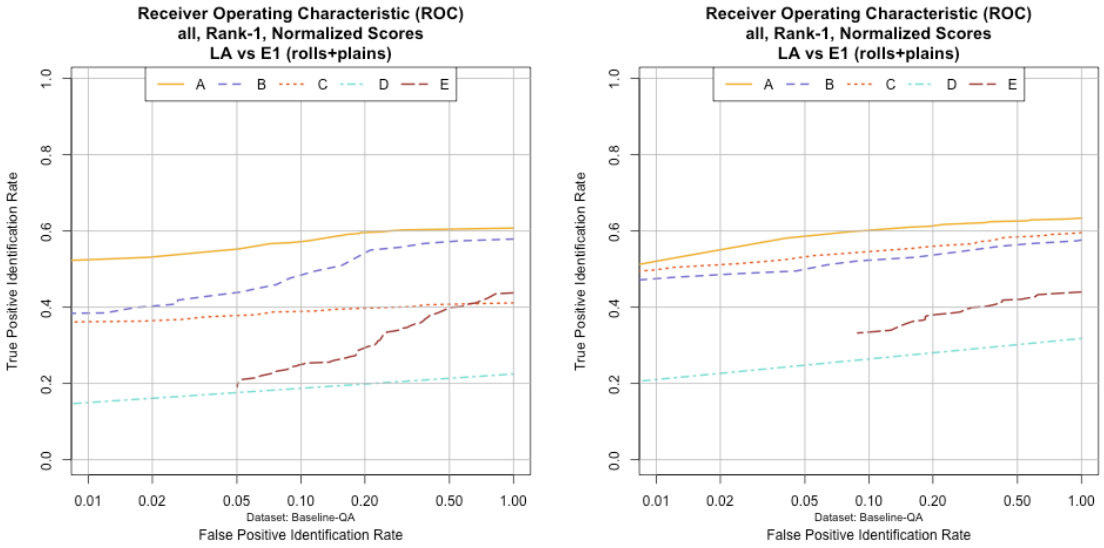

Figure 3A Performance Comparison, LA vs. E1 - Evaluation #1 on left, #2 on right

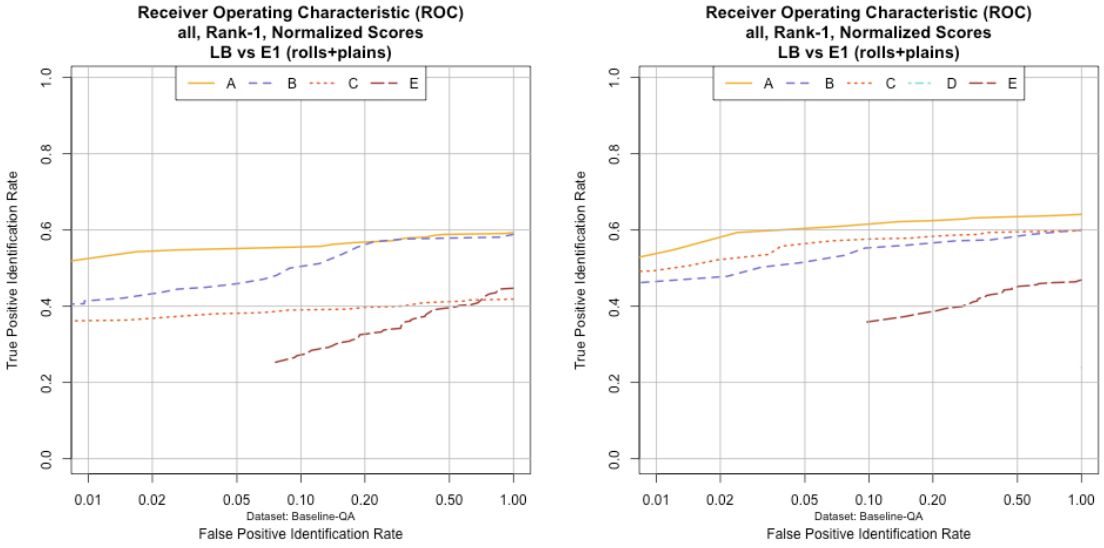

Figure 3B: Performance Comparison, LB vs. E1 - Evaluation #1 on left, #2 on right

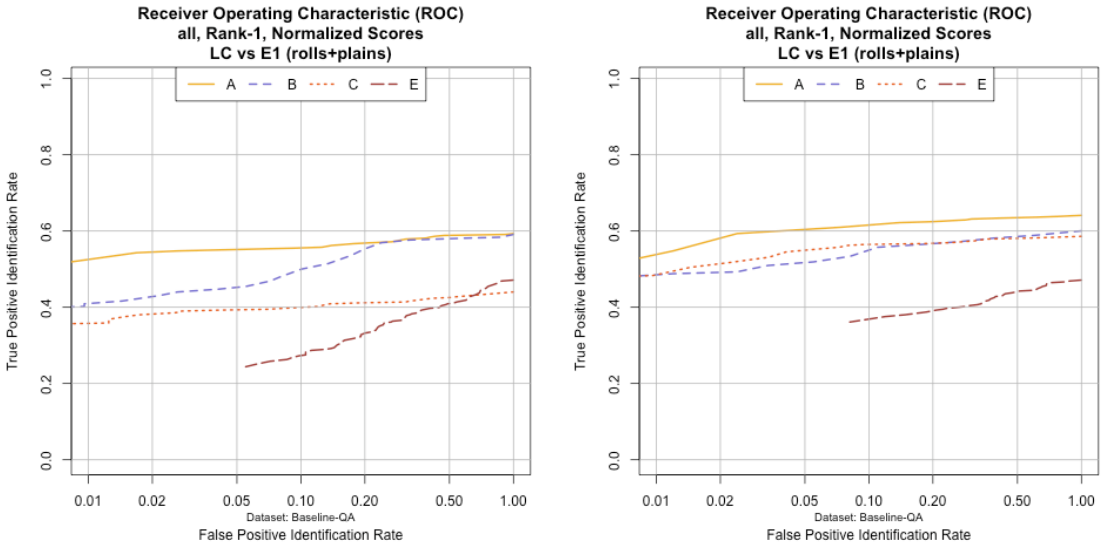

Figure 3C: Performance Comparison, LC vs. E1 - Evaluation #1 on left, #2 on right

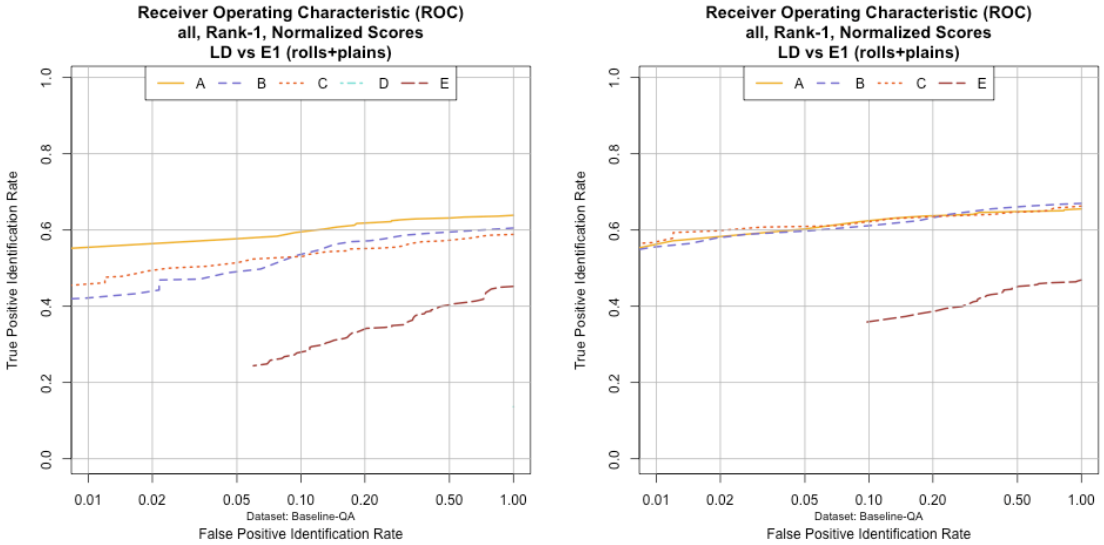

Figure 3D: Performance Comparison, LD vs. E1 - Evaluation #1 on left, #2 on right

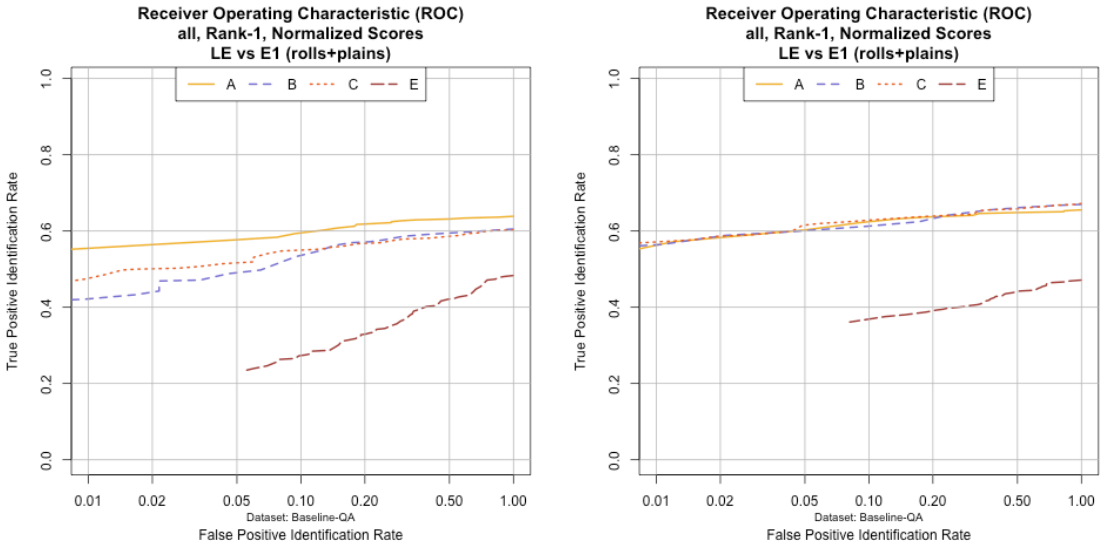

Figure 3E: Performance Comparison, LE vs. E1 - Evaluation #1 on left, #2 on right

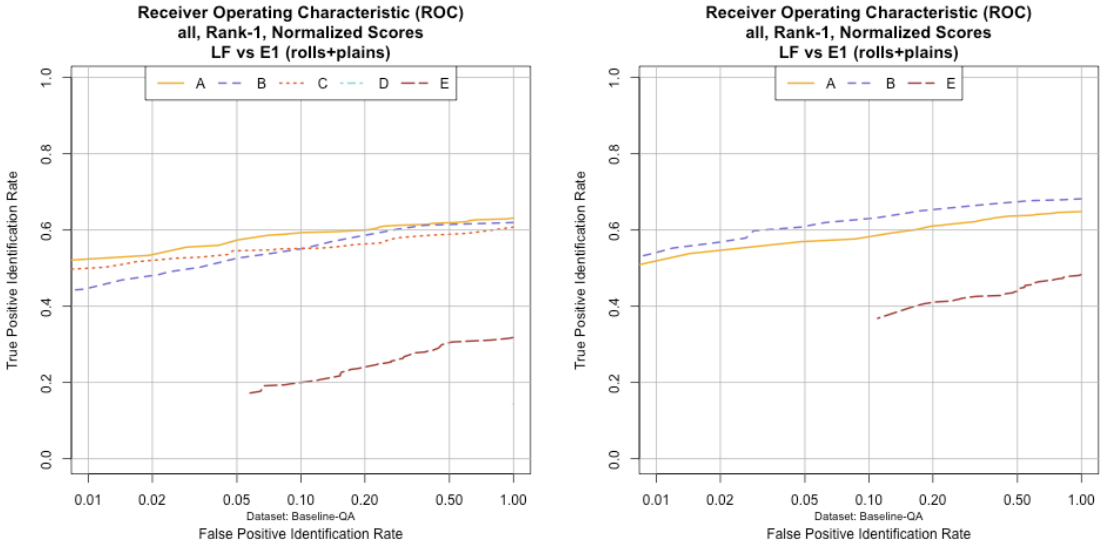

Figure 3F: Performance Comparison, LF vs. E1 - Evaluation #1 on left, #2 on right

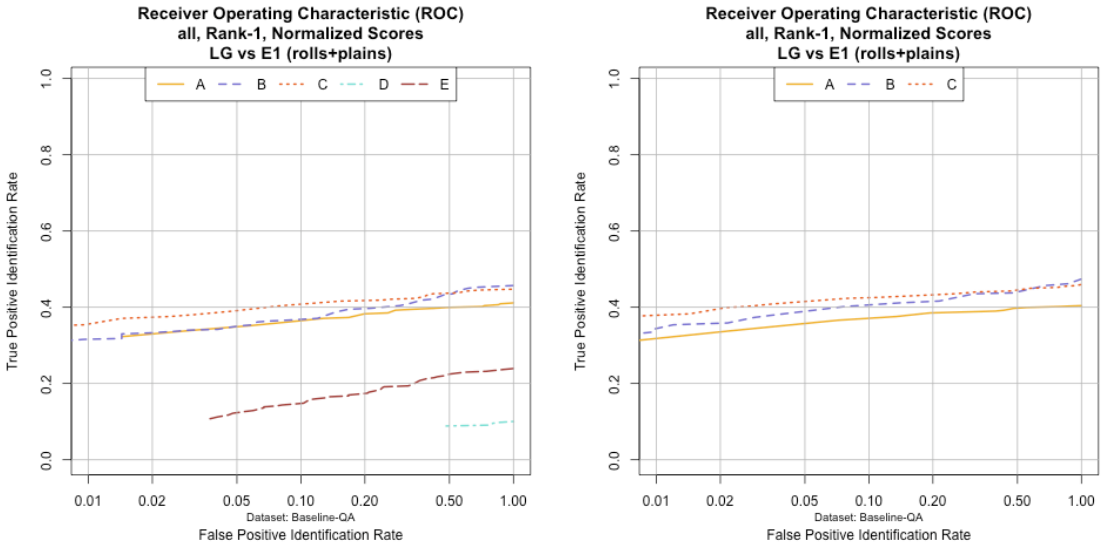

Figure 3G: Performance Comparison, LG vs. E1 - Evaluation #1 on left, #2 on right

Figure 3 (parts A-G): Score-based comparison of matchers for latent subsets LA-LG

(Baseline-QA dataset, 418 latents — 100,000 rolled+plain 10-finger exemplar sets)

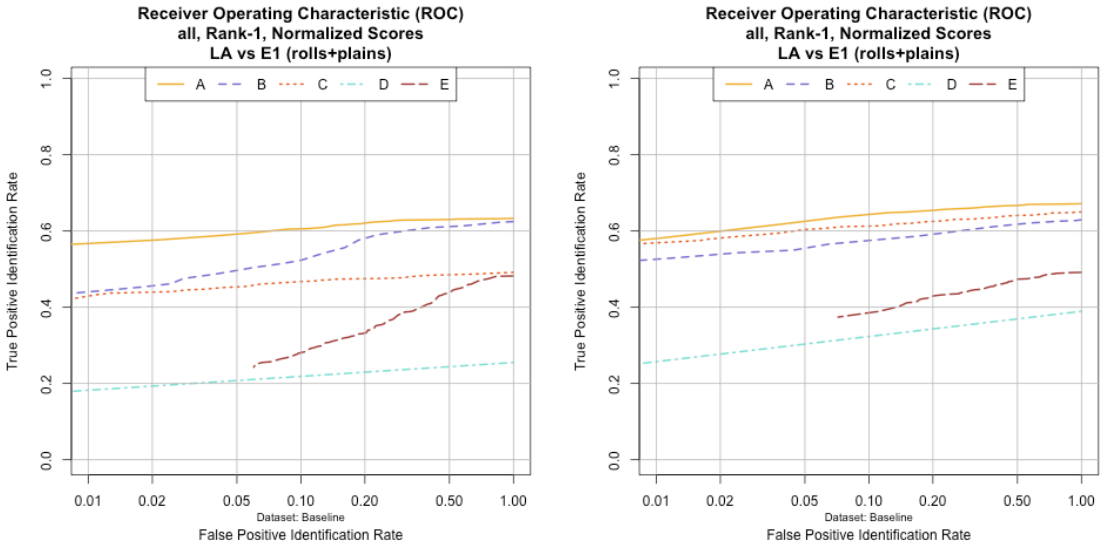

Figure 4A: Performance Comparison, LA vs. E1 - Evaluation #1 on left, #2 on right

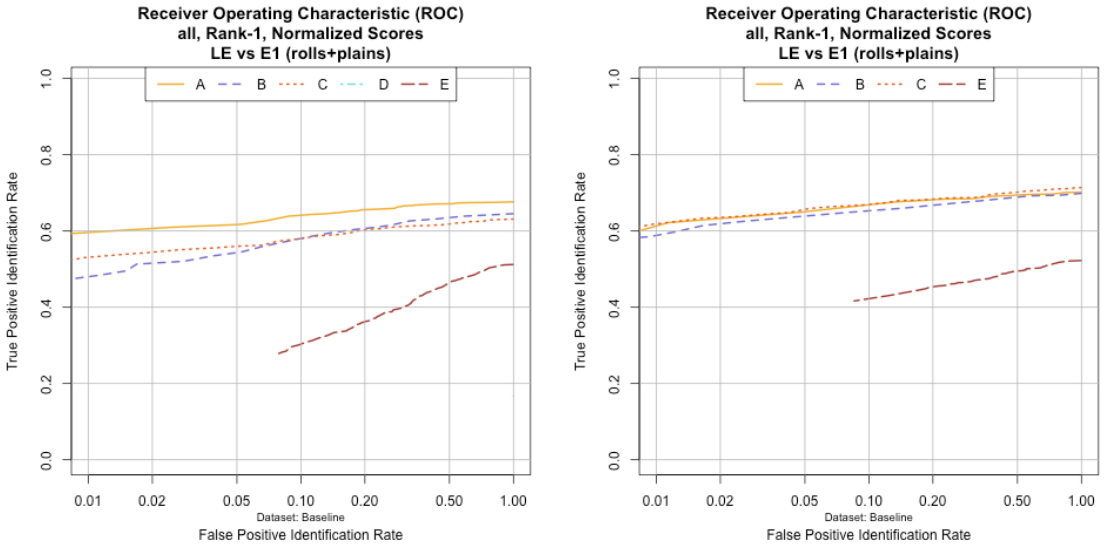

Figure 4B: Performance Comparison, LE vs. E1 - Evaluation #1 on left, #2 on right

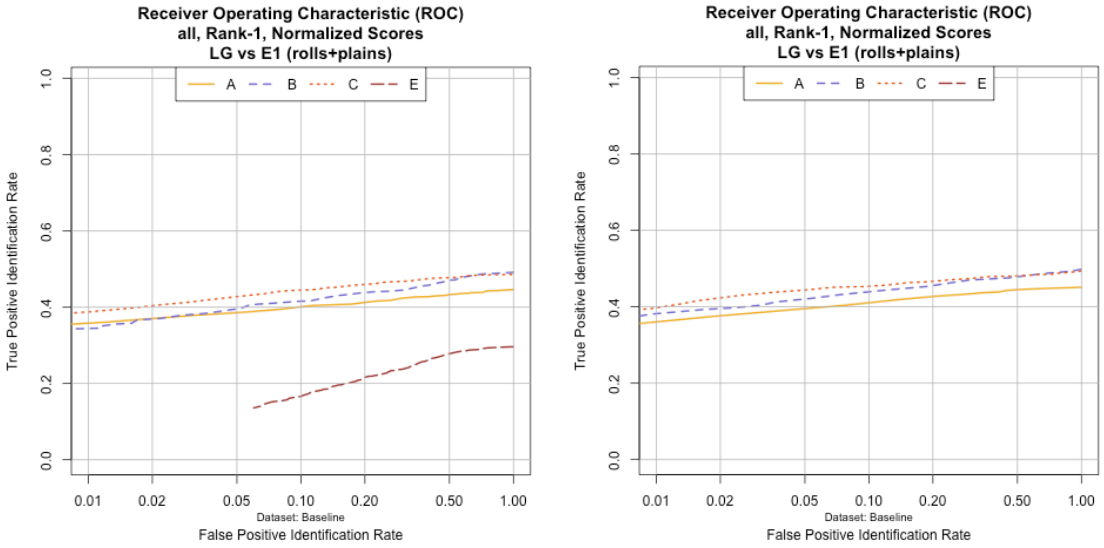

Figure 4C: Performance Comparison, LG vs. E1 - Evaluation #1 on left, #2 on right

Figure 4 (parts A,B,C): Score-based comparison of matchers for latent subsets LA,LE,LG

(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

The order of the matchers with respect to accuracy varies depending on FPIR: while the CMC curves were generally parallel, the ROC curves often cross. Note that two matchers can have similar performance at a high FPIR, but show substantial differences in accuracy as FPIR approaches 0.01. For example, compare matchers B and C in subsets LB, LC, and LG, or matchers A, B, and C in subsets LD and LE.

For Evaluation 2, at low values of FPIR (0.01), the highest overall accuracy was by matcher C using the LD feature subset for Baseline-QA, and the LE feature subset for Baseline. Matcher A had the best accuracy for LA-LC based searches; Matcher C had the best accuracy for LD-LE and LG based searches; Matcher B had the highest performance for LF based searches.

The following tables summarize the Evaluation 2 results from the ROCs above. Table 11 shows the TPIR (rank 1, normalized score) at an FPIR of 0.01 for the Baseline-QA dataset; Table 12 shows the corresponding results for the

Baseline dataset. For cases in which the TPIR could not be measured at FPIR=0.01 for the specified matcher-subset combination, the closest point is indicated in gray. Highlights follow the pattern used in Table 9 and Table 10.

Table 11: TPIR at FPIR=0.01, for all Latent Feature Subsets

(Baseline-QA dataset, 418 latents — 100,000 rolled+plain 10-finger exemplar sets)

Table 11A – ELFS-EFS Evaluation #1 results

		Latent Feature Subset						
		LA	**LB**	**LC**	**LD**	**LE**	**LF**	**LG**
Matcher	**A**	52.2	52.9	52.9	55.0	55.0	51.9	32.3
	B	38.5	41.4	40.9	42.1	42.1	44.5	31.6
	C	36.1	36.4	35.9	47.6	48.6	49.5	35.4
	D	17.0	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*9.8@ FPIR=47.6*
	E	*19.1@ FPIR=0.05*	*25.1@ FPIR=0.07*	*24.4@ FPIR=0.05*	*24.4@ FPIR=0.06*	*23.4@ FPIR=0.06*	*17.2@ FPIR=0.06*	*10.5@ FPIR=0.04*

Table 11B – ELFS-EFS Evaluation #2 results

		Latent Feature Subset						
		LA	**LB**	**LC**	**LD**	**LE**	**LF**	**LG**
Matcher	**A**	52.9	54.8	54.8	57.2	57.2	50.8	30.9
	B	47.8	46.9	48.8	56.5	56.5	53.8	35.4
	C	50.5	50.7	50.5	57.9	57.6	*n/a*	38.3
	D	21.0	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
	E	*33.3@ FPIR=0.09*	*35.9@ FPIR=0.10*	*36.1@ FPIR=0.08*	*35.9@ FPIR=0.10*	*36.1@ FPIR=0.08*	*36.8@ FPIR=0.11*	*n/a*

Table 12: TPIR at FPIR=0.01, for Latent Feature Subsets LA, LE, LG

(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

Table 12A – ELFS-EFS Evaluation #1 results

		Latent Feature Subset		
		LA	**LE**	**LG**
Matcher	**A**	56.4	59.4	36.1
	B	43.4	47.1	34.3
	C	42.4	53.0	38.8
	D	18.0	*n/a*	*n/a*
	E	*24.3 @ FPIR=0.06*	*27.9 @ FPIR=0.08*	*13.6 @ FPIR=0.06*

Table 12B – ELFS-EFS Evaluation #2 results

		Latent Feature Subset		
		LA	**LE**	**LG**
Matcher	**A**	58.3	62.2	37.5
	B	53.1	59.8	38.5
	C	57.5	62.5	40.4
	D	24.8	*n/a*	*n/a*
	E	*37.4 @ FPIR=0.07*	*41.7 @ FPIR=0.08*	*n/a*

7 Effect of Rolled and/or Plain Exemplars

The results discussed so far have been based on searches of gallery E1 (described in Table 8) which consists of ten rolled and ten plain impressions for each subject. Searches were also conducted against two other galleries, E2 and E3 which consist, respectively, of only ten rolled impressions per subject, and only ten plain impressions per subject, in order to compare the rank-1 accuracy for the different exemplar types. A general expectation would be that rolled+plain (E1) accuracy would always be higher than either rolled (E2) or plain (E3) separately, and that accuracy for rolled (E2) would generally be higher than plain (E3). While Evaluation #1 indicated that these expectations were generally but not always true, Evaluation #2 showed these expectations were true without exception.

The following tables compare the Evaluation #1 and Evaluation #2 results. Cases which did not conform to expectations are highlighted in yellow (i.e. rolled performance was worse than plain, or rolled+plain was worse than either rolled or plain).

Table 13: Rank-1 accuracy by exemplar type: rolled+plain (E1); rolled (E2); plain (E3)
(Baseline dataset, 1066 latents —10-finger exemplar sets: E1: 100,000 rolled+plain; E2: 10,000 rolled; E3: 10,000 plain)

Table 13A – ELFS-EFS Evaluation #1 results

		Latent Feature Subset / Exemplar Type								
		LA			LE			LG		
		E1	E2	E3	E1	E2	E3	E1	E2	E3
Matcher	A	63.3	58.0	50.8	67.7	61.8	53.1	44.7	40.4	33.2
	B	62.5	58.2	50.4	64.5	62.1	53.9	49.2	56.9	48.7
	C	49.2	47.7	41.8	63.1	58.3	50.1	48.6	44.1	45.6
	D	25.5	19.7	20.5	16.7	14.5	11.6	11.7	11.9	11.5
	E	48.2	41.8	36.7	51.2	45.0	38.8	29.6	25.7	21.4

Table 13B – ELFS-EFS Evaluation #2 results

		Latent Feature Subset / Exemplar Type								
		LA			LE			LG		
		E1	E2	E3	E1	E2	E3	E1	E2	E3
Matcher	A	67.2	62.3	33.1	70.2	64.8	34.2	45.1	41.4	33.3
	B	63.0	60.1	52.2	69.9	67.3	58.6	49.8	48.4	40.5
	C	65.0	58.4	52.3	71.4	63.0	55.2	49.3	44.6	37.2
	D	38.9	34.7	29.0	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
	E	49.2	43.1	38.6	52.3	46.1	39.9	0	0.1	0.1

Note that the performance inversions in Evaluation #1 have disappeared in Evaluation #2. Table 14 shows the data from Table 13 in terms of relative gains in accuracy. In all cases rolled+plain is more accurate than rolled alone (E1-E2), and is much more accurate than plain alone (E1-E3); in all cases, rolled alone is more accurate than plain alone (E2-E3). However, matcher A showed an unusually large drop-off in performance for plain impressions in comparison to the other matchers (for latent subsets LA and LE).

Table 14: Differences in rank-1 accuracy by exemplar type: rolled+plain vs. rolled (E1-E2); rolled+plain vs. plain (E1-E3); rolled vs. plain (E2-E3)
(Baseline dataset, 1066 latents —10-finger exemplar sets: E1: 100,000 rolled+plain; E2: 10,000 rolled; E3: 10,000 plain)

Table 14A – ELFS-EFS Evaluation #1 results

		E1-E2			E1-E3			E2-E3		
		LA	LE	LG	LA	LE	LG	LA	LE	LG
Matcher	A	5.3	5.9	4.3	12.5	14.6	11.5	7.2	8.7	7.2
	B	4.3	2.4	-7.7	12.1	10.6	0.5	7.8	8.2	8.2
	C	1.5	4.8	4.5	7.4	13.0	3.0	5.9	8.2	-1.5
	D	5.8	2.2	-0.2	5.0	5.1	0.2	-0.8	2.9	0.4
	E	6.4	6.2	3.9	11.5	12.4	8.2	5.1	6.2	4.3

Table 14B – ELFS-EFS Evaluation #2 results

		E1-E2			E1-E3			E2-E3		
		LA	LE	LG	LA	LE	LG	LA	LE	LG
Matcher	A	4.9	5.4	3.7	34.1	36.0	11.8	29.2	30.6	8.1
	B	2.9	2.6	1.4	10.8	11.3	9.3	7.9	8.7	7.9
	C	6.6	8.4	4.6	12.7	16.2	12.1	6.1	7.8	7.4
	D	4.2	*n/a*	*n/a*	9.9	*n/a*	*n/a*	*5.7*	*n/a*	*n/a*
	E	6.1	6.2	-0.1	10.6	12.4	-0.1	4.5	6.2	0

Observations

- The greatest effect is the difference between matching against rolled vs. plain. Use of rolls results in an average performance gain of approximately 11 percentage points over use of plains.
- For comparison the difference between E1 (rolled + plain) and E2 (rolled only) is less than half of this (4.4 percentage points). Note that this difference is much smaller for data type LG than the other two. This can be partially, but not entirely, explained by the fact that performance is lower for this data type; but some unexplained factor appears to be present.
- Matcher A showed a very large drop-off in performance for latent subsets LA and LE when plain impressions were used. This drop-off was not consistent with the other matchers.

8 Effect of Examiner Markup Method

The results discussed so far in this report have been based on human examiner markup of each latent image, as might be expected in casework. This approach has the benefit of being realistic, but for the purposes of algorithmic performance evaluation there is a drawback in that the latent markup includes the variability one might expect from any human activity, including the results of differences of expertise and possible error. When evaluating matchers, one approach taken in the past was to mark only "Ground Truth" features, by referring to the exemplar(s) when marking each latent, so that the result would include no false or missed features. This groundtruthing process obviously cannot be used operationally, but it has been used very effectively to provide idealized test datasets that minimize human variability for the purposes of development and evaluation of fingerprint feature extraction and matching software.

In every other section of this report, all latent images were marked by examining the latent in isolation. The results discussed in this section show the differences between operationally practical markup and groundtruth markup. The set of latent images in Baseline-QA were marked by the examiners using three different approaches, as defined in Table 15.

Table 15: Methods of examiner markup

Markup method	Description
(standard)	Features were marked in latent images without reference to exemplars. Used for Baseline and Baseline-QA.
AFIS	Derived from the Baseline-QA markup. The examiners removed debatable minutiae, to test the assumption that false minutiae are worse than missed minutiae for AFIS searching.[h]
GT	Ground Truth markup, derived from the Baseline-QA markup. Exemplar images (both rolled and plain) were consulted when marking latent features, so that all latent minutiae[i] marked were corroborated by one or more exemplars. Note that this is not possible operationally, but defines an upper bound for the accuracy of feature matching.

These results depict the difference in accuracy between theoretically ideal markup and operational human markup. Table 16 and Table 17 summarize rank-1 performance data for searches using the different feature markup sets. In each case, the same set of latents is used, varying only by the markup. Groundtruth is indicated with a "G" superscript, and AFIS is indicated with an "A" superscript (e.g. LE^G, or LE^A).

Table 16: Comparison of rank-1 IR for standard, groundtruth, and AFIS markup methods (LE and LG)
(Baseline-QA dataset, 418 latents — 100,000 rolled+plain 10-finger exemplar sets)

Table 16A – ELFS-EFS Evaluation #1 results

	LA		LE	LE^G	LE^A		LG	LG^G	LG^A
A	61.8		64.7	69.6	63.5		42.4	53.4	41.7
B	59.1		62.0	69.6	61.8		47.1	62.0	48.3
C	41.7		60.8	56.4	49.3		45.3	60.0	47.1
D	22.8		15.2	15.5	16.2		10.3	12.8	8.8
E	44.6		49.8	48.5	49.3		24.5	37.2	23.0

Table 16B – ELFS-EFS Evaluation #2 results

	LA		LE	LE^G	LE^A		LG	LG^G	LG^A
A	64.5		66.7	71.8	60.8		41.7	53.7	38.5
B	58.8		68.4	74.3	63.0		48.8	61.0	44.4
C	60.1		67.9	72.3	62.8		46.6	61.5	44.4
D	32.4		n/a	n/a	n/a		n/a	n/a	n/a
E	45.1		48.8	48.5	44.1		0.0	0.0	0.0

Table 17: Differences in rank-1 IR for standard, groundtruth, and AFIS markup methods (LE and LG)
(Baseline-QA dataset, 418 latents — 100,000 rolled+plain 10-finger exemplar sets)

Table 17A – ELFS-EFS Evaluation #1 results

	Effect of Ground truth markup		**Effect of AFIS markup**	
	LE^G - LE	LG^G – LG	LE^A – LE	LG^A-LG
A	4.9	11.0	-1.2	-0.7
B	7.6	14.9	-0.2	1.2
C	-4.4	14.7	-11.5	1.8
D	0.3	2.5	1.0	-1.5
E	-1.3	12.7	-0.5	-1.5

[h] *A review was conducted to derive a generic set of rules for AFIS feature markup, considering possible rules such as excluding minutiae on short ridges or short enclosures, excluding minutiae near the core or in high curvature areas, excluding isolated minutiae, or excluding separated clusters of minutiae. Each of these potential changes was determined not to be vendor neutral, and was rejected because it had the potential to benefit some participants to the detriment of others. The resulting guidance was solely to remove the most debatable minutiae."*

[i] *Note that the minutiae were groundtruthed against the exemplars, but not the other extended features, such as incipients or skeletons.*

Table 17B – ELFS-EFS Evaluation #2 results

	Effect of Ground truth markup		Effect of AFIS markup	
	LE^G - LE	LG^G – LG	LE^A – LE	LG^A-LG
A	5.1	12.0	-5.9	-3.2
B	5.9	12.2	-5.4	-4.4
C	4.4	14.9	-5.1	-2.2
D	n/a	n/a	n/a	n/a
E	-0.3	0.0	-4.7	0.0

The groundtruth (GT) results were beneficial using latent subset LE (matchers A, B, and C), but were dramatically beneficial for using latent subset LG (matchers A, B, and C). In practice this means that for minutiae-only markup (LG), there is about 12-15% difference in rank-1 accuracy between ideal and ordinary examiner markup; this difference is pronounced since the matcher had no recourse to the image. When the image and features are included (LE), the difference drops to about 4-6%. Matchers B and C derived the most benefit overall from "GT". markup. Note that for matcher A, image-only matching is more accurate than minutiae-only matching even when using groundtruth features.

The "AFIS" markup approach provided was counterproductive in all cases. This result indicates that the matchers are relatively robust when processing debatable minutiae.

.

9 Effect of Latent Data Source

The latents were collected from disparate sources. The following charts show the difference in rank-1 identification rate between the sources of latents for the Baseline dataset. As shown above in Table 2, Casework 1 and 2 were from operational casework, while the others were collected in laboratory conditions.

All matchers were shown to be sensitive to dataset characteristics that are created as part of the dataset capture process. The difference in rank-1 IR between the WVU and the FLDS sources was over 20% for all matchers A, D and E using latent feature subset LA, and was even higher for participants B and C. For latent feature subset LE, a similar difference was noted for participants A, B, C, and E. The differences in rank-1 identification rate across the data sources were even greater when latent feature subset LG was used. The WVU data, which contains the greatest average number of minutiae, was approximately 40% more accurate for matchers A and C, and nearly 30% more accurate for matcher B when compared with Casework 2.

Nearly all matchers improved their Evaluation #1 performance with respect to data source. A notable exception was matcher B whose performance dropped on the Casework 2 and FLDS datasets for subset LA.

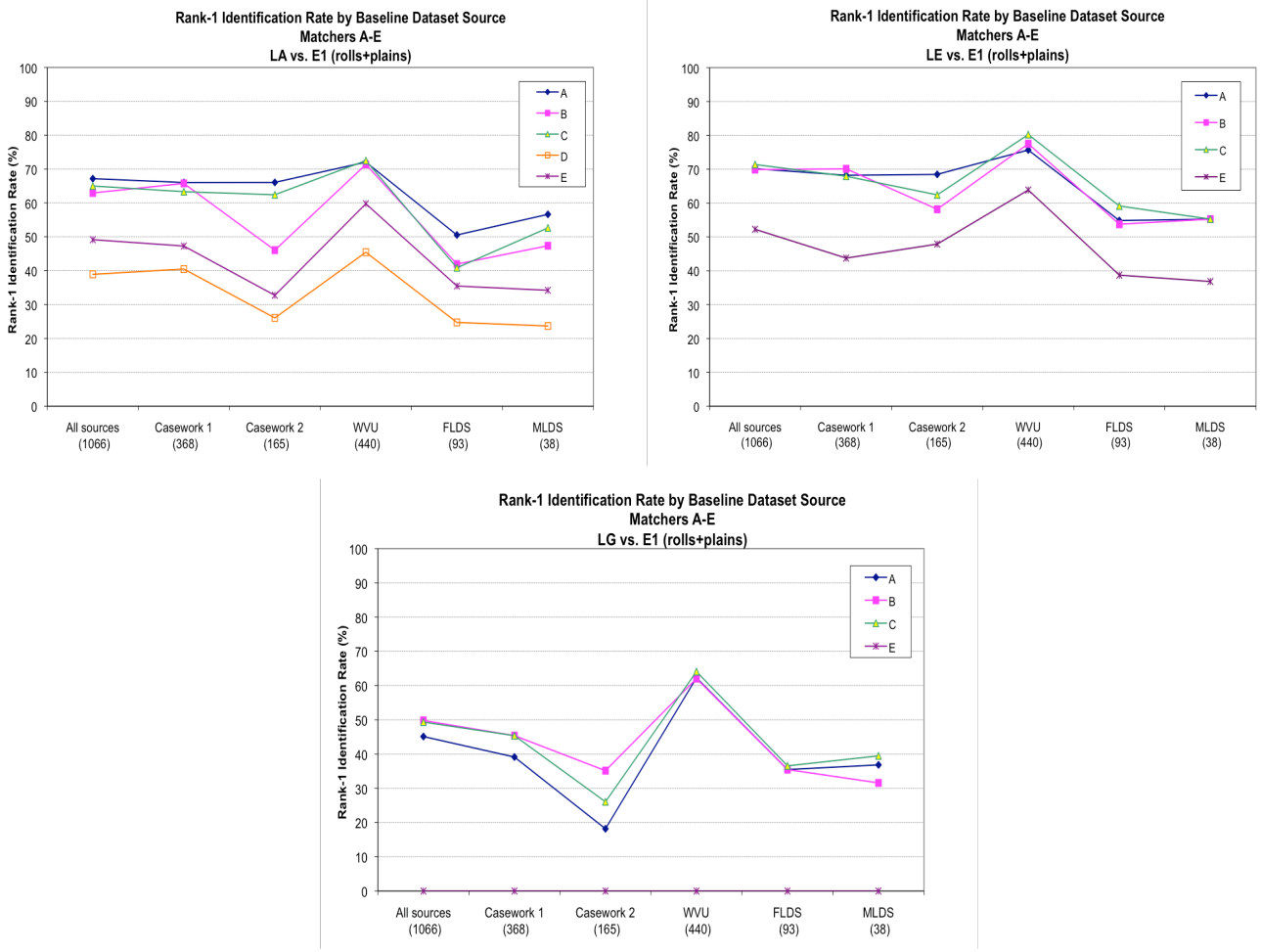

Figure 5: Comparison of rank-1 IR by latent data source
(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

10 Effect of Latent Orientation

This analysis shows the effect of orientation on rank-1 identification rate (see Section 4.1.4). The latents from the Baseline dataset were grouped by orientation (rotation from upright). Note that the first two bins are aggregates of the other bins: 'All' contains all latents in the Baseline dataset, regardless of orientation; '0-45 degrees' is shown because it illustrates a range typical for operational searches. The latents in bin '0-45 degrees' comprise 87.4 % of the latents in the Baseline dataset. LE and LG searches included orientation information (when known); for LA searches, the orientation was never known to the participants.

Nearly all matchers improved their Evaluation #1 performance for nearly every range of orientation. A notable improvement was for matcher B, subset LA, for '90-180 degrees' which improved its identification rate from 0 to 46% .

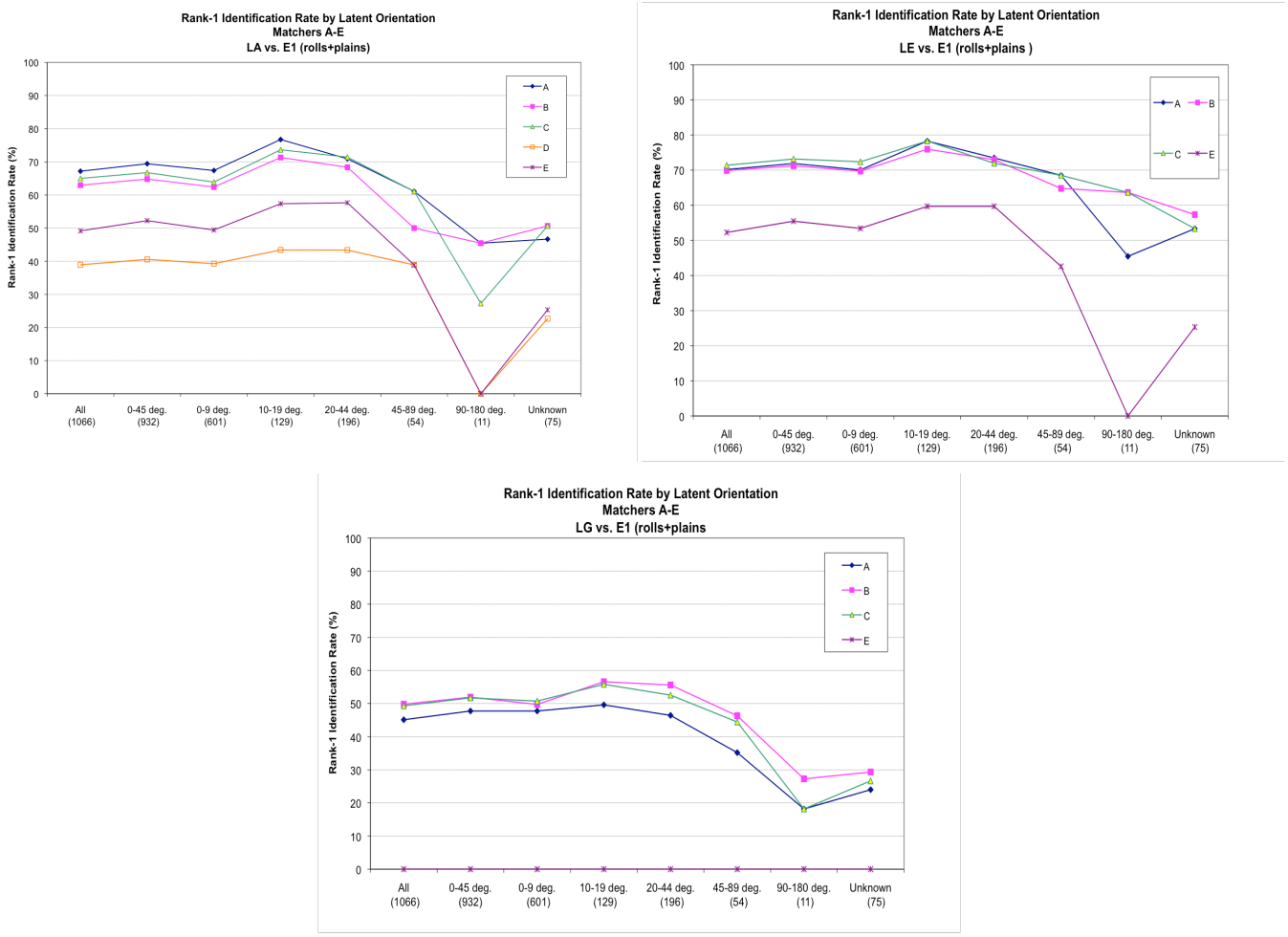

Figure 6: Comparison of rank-1 IR by orientation of latent image
(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

For all matchers on all latent feature subsets, searches of latents with a known orientation in the range 0 to 45 degrees are on average more accurate than for latents that are 45-90 degrees from vertical, and are usually much more accurate than latents rotated more than 90 degrees.

11 Effect of Latent Minutiae Count

The following charts show rank-1 identification rates broken into bins by minutiae count, for latent subsets LA, LE, and LG. See Appendix A for the Evaluation #1 results on this set. True positive identification rates (TPIR) at rank-1 and FPIR=0.01, where available, are also shown for comparison.

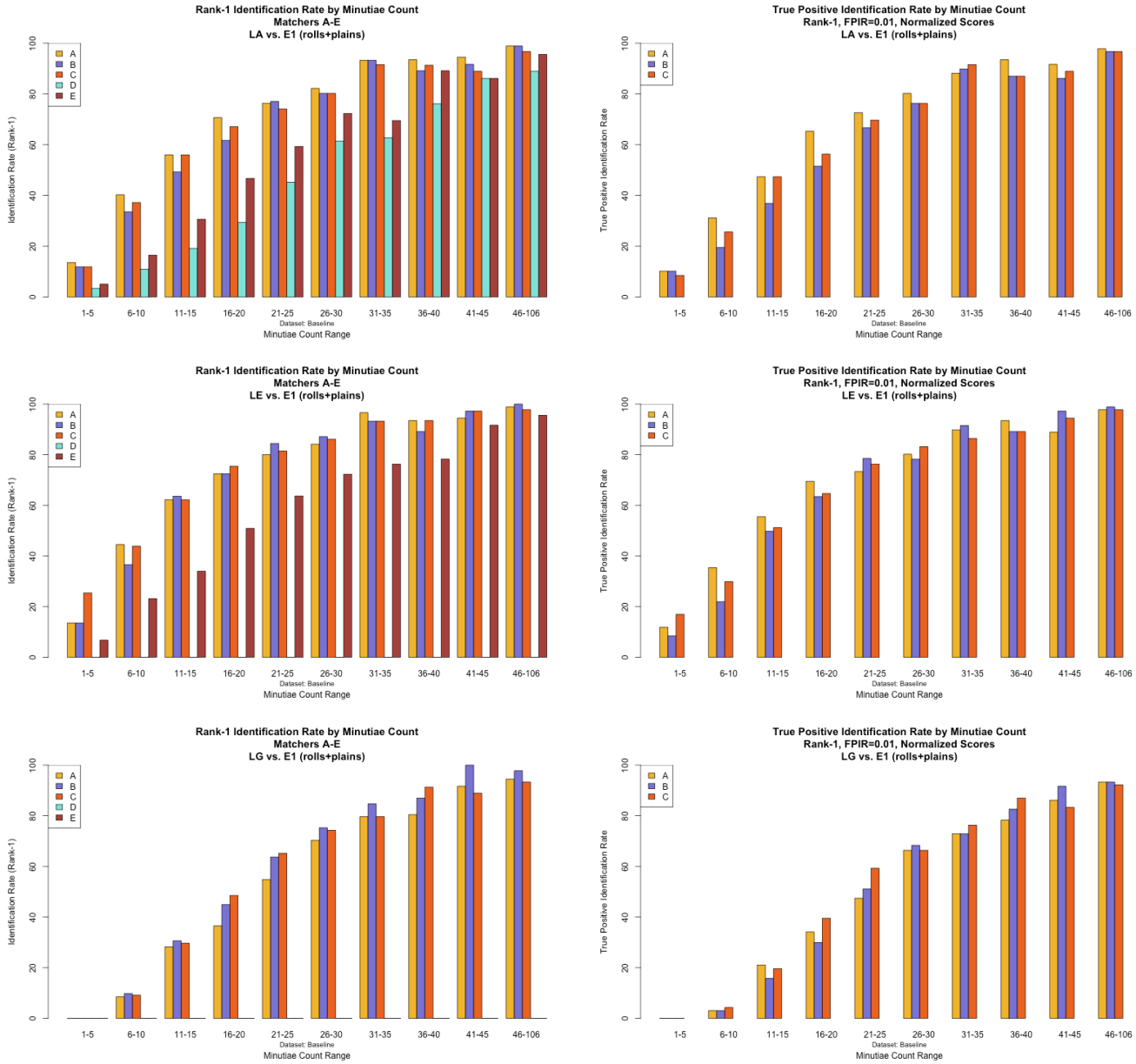

Figure 7: Rank and score-based accuracy by number of minutiae
(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

The following chart shows rank-1 identification rate change (difference) for searches of latents in the Baseline dataset, between latent subset LA (image-only) and latent subset LE (image+EFS), broken out by minutiae count range. A positive difference indicates an increase in rank-1 identification rate performance when image + EFS (LE) searches are used instead of image-only (LA) searches.

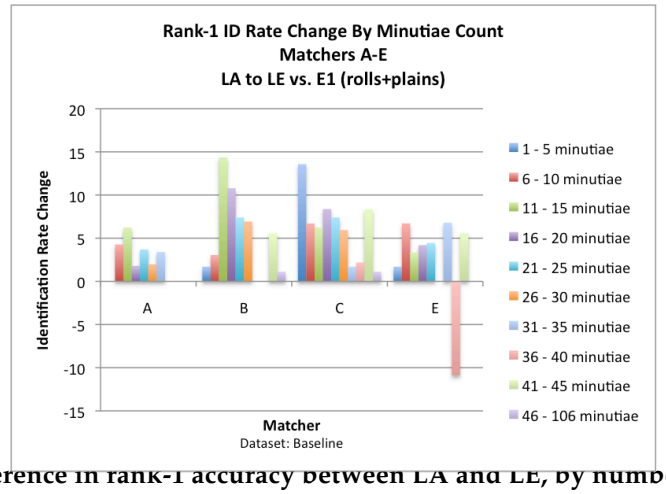

Figure 8: Difference in rank-1 accuracy between LA and LE, by number of minutiae
(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

Performance improves linearly as a function of minutiae count, however, the performance levels off significantly for all matchers at about 30 minutiae.

Matchers A,B and C achieve a 12-14% identification rate even on latents with 1-5 minutiae, for subset LA; matcher A's corresponding TPIR drops to 10% at FPIR=0.01.

Nearly all matchers improved their Evaluation #1 performance (see Appendix A for the Evaluation #1 results on this set) for nearly every range of minutiae count.

Most matchers showed performance increases for image+feature searches (LE or LD) compared with image-only (LA) searches across all minutiae count ranges. For most matchers this performance increase grew in magnitude as minutiae count increased, before peaking and then dropping.

12 Effect of Latent Value Determination

A discussed in Section 4, the examiners who marked the latent images made determinations of Value, Limited Value (Value for exclusion only), or No Value at the time of markup (see Table 6 for minutiae counts by value determination). Table 18 and Table 19 show the relationship between value determination and rank-based and score-based accuracy. Value determination assessments are only in reference to the latent, and do not consider exemplar quality.

Table 18: Rank-1 identification rate by value determination
(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

		All	No Value	Limited Value	Value
Count		1066[j]	25	113	917
LA	A	67.2%	20.0%	34.5%	72.6%
	B	63.0%	8.0%	28.3%	68.4%
	C	65.0%	8.0%	30.1%	70.8%
	D	38.9%	4.0%	4.4%	44.1%
	E	49.2%	0.0%	10.6%	55.0%
LE	A	70.2%	20.0%	35.4%	75.9%
	B	69.9%	12.0%	31.0%	76.2%
	C	71.4%	20.0%	35.4%	77.1%
	D	*n/a*	*n/a*	*n/a*	*n/a*
	E	52.3%	0.0%	17.7%	57.9%
LG	A	45.1%	4.0%	6.2%	51.2%
	B	49.8%	0.0%	4.4%	56.8%
	C	49.3%	0.0%	7.1%	55.8%
	D	*n/a*	*n/a*	*n/a*	*n/a*
	E	0.0%	0.0%	0.0%	0.0%

Table 19: Rank-1 TPIR at FPIR=1% by value determination
(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

		All	No Value	Limited Value	Value
Count		1066	25	113	917
LA	A	58.3%	12.0%	28.3%	67.5%
	B	53.1%	4.0%	19.5%	59.9%
	C	57.5%	4.0%	21.2%	64.5%
	D	24.8%	0.0%	0.0%	17.1%
	E	-	-	-	-
LE	A	62.2%	16.0%	30.9%	70.8%
	B	59.8%	12.0%	20.4%	67.9%
	C	62.5%	16.0%	25.7%	69.0%
	D	-	-	-	-
	E	-	-	-	-
LG	A	37.5%	4.0%	1.8%	46.2%
	B	38.5%	0.0%	2.7%	45.6%
	C	40.4%	0.0%	2.7%	49.3%
	D	-	-	-	-
	E	-	-	-	-

Figure 9 shows rank-1 identification rate change (difference) for searches of latents in the Baseline-QA dataset, between various latent subsets, broken out by latent value determination. For example in the first chart below, "LA to LD" means that the first set of searches used the feature subset LA, and the second set used the feature subset LD, thus the difference in ID rate is a measure of "benefit" from supplying minutiae and ridge count information to the matcher in addition to the image. In some latent value determination classes the difference in ID rate is not always positive: a negative ID rate differences indicates that the matcher actually performed worse when additional features were added.

[j] *Note: 11 latents (out of 1066) in Baseline, and 5 latents (out of 418) in Baseline-QA did not have value determinations.*

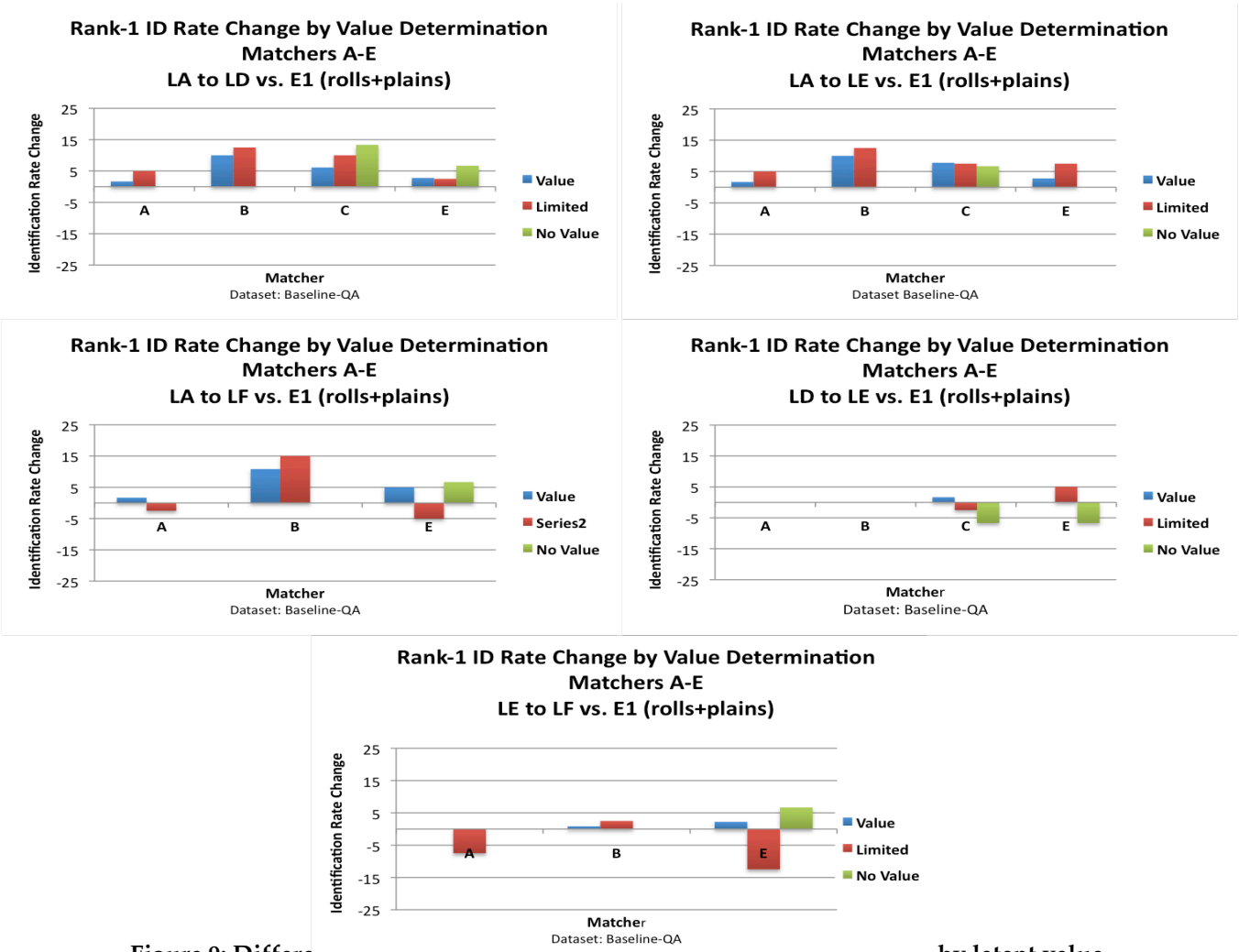

Figure 9: Difference in rank-1 identification rate between latent subsets, by latent value
(Baseline-QA dataset, 418 latents — 100,000 rolled+plain 10-finger exemplar sets)

As expected, accuracy is very clearly related to latent value determination, with much greater accuracy for the latents determined a priori to be of value. A notable and surprising result is that some fingerprints assessed as being of No or Limited Value were capable of being identified by the matchers being tested. In particular, matcher A's rank-1 identification rate for No Value latents is 20% on subsets LA/LE, and even higher (35%) on Limited Value latents. Their corresponding TPIR rates decrease at FPIR=0.01, though they are still notable. Matcher C also had a rank-1 identification rate for No Value latents of 20% for LE, and 35% for Limited Value latents. These results echo the performance on latents with few minutiae, as discussed in Section 11. The results for participants A, B, C, and E show that matching may be practical even for Limited Value or No Value latents, given lower expectations of accuracy (latents considered to have no value may not be subject to individualization or elimination, but they could be used as an investigative tool). The decision to include No Value and Limited Value latents in the test has been justified. The ability to match these low value prints requires the image; none of the No-Value images could be matched using minutiae alone at rank 1 or FPIR=0.01.

The ordering of matchers by rank-1 identification rate tended to remain the same across different latent value determinations. Note that Matchers A and C have equivalent performance for Limited Value latents with subset LE.

Nearly all matchers improved their Evaluation #1 performance for nearly every category of latent value determination. Matchers B and C improved the most for Limited Value and No Value latents.

13 Effect of Latent Good / Bad / Ugly Quality Classifications

As discussed in Section 4, a set of latent examiners categorized the latents according to an subjective Excellent, Good, Bad, Ugly, Unusable quality scale[k]; these examiners were different from those providing feature markup or value assessments. See Table 6 for minutiae counts by value determination; see Table 5 for comparison of value and quality assessments.

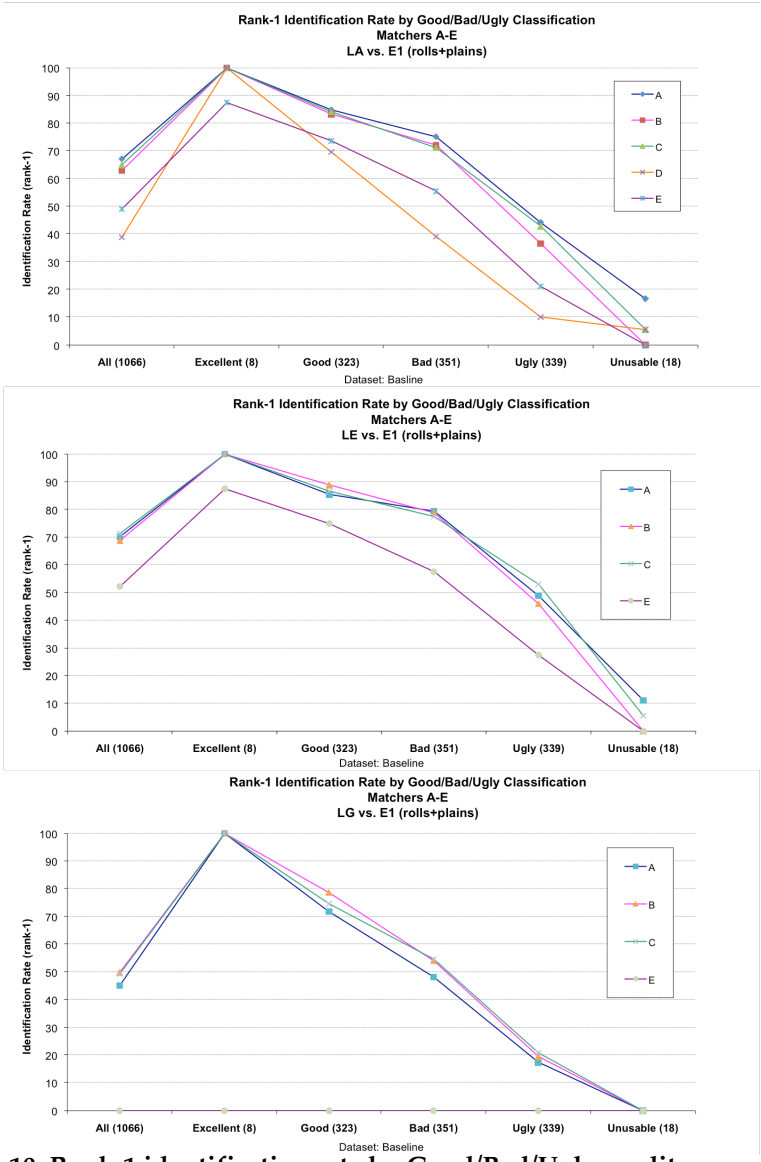

Figure 10: Rank-1 identification rate by Good/Bad/Ugly quality assessment
(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

The following charts show rank-1 identification rate change (difference) for latents in the Baseline-QA dataset between different searches of latent feature subsets, broken out by Good / Bad / Ugly classification. A positive difference indicates an increase in rank-1 ID rate performance between successive sets of searches of the same latents where the second set of searches supplies additional feature information to the matcher. For example in the first chart below, "LA to LD" means that the first set of searches used the feature subset LA, and the second set used the feature subset LD, thus the difference in ID rate is a measure of "benefit" from supplying minutiae and ridge count information to the matcher in addition to the image. In some Good / Bad / Ugly classes the difference

[k] *27 of the Baseline latents, including 9 of the Baseline-QA latents, did not have the Good-Bad-Ugly quality assessments.*

in ID rate is not always positive, a negative ID rate differences indicates that the matcher actually performed worse when additional features were added.

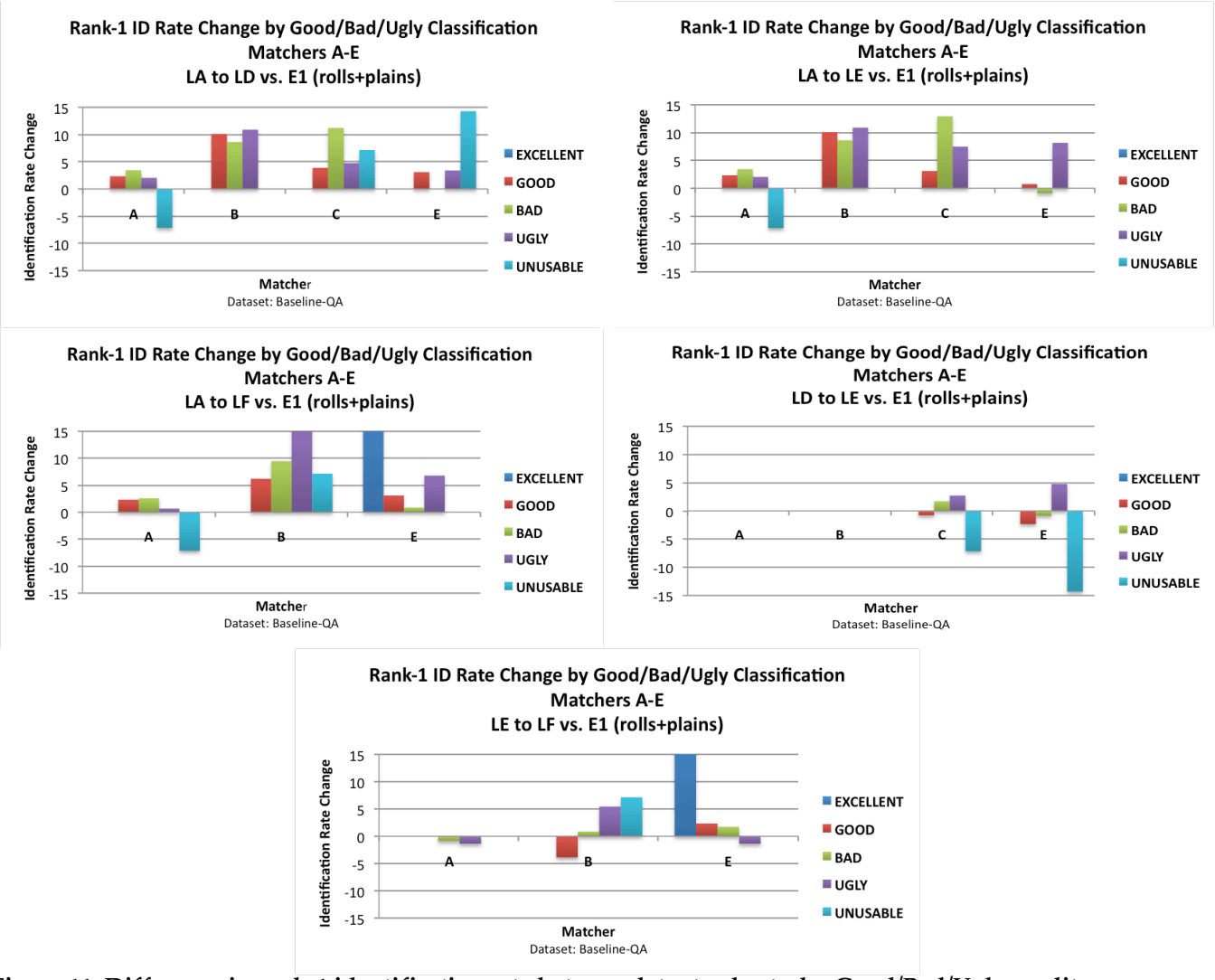

Figure 11: Difference in rank-1 identification rate between latent subsets, by Good/Bad/Ugly quality assessment
(Baseline-QA dataset, 418 latents — 100,000 rolled+plain 10-finger exemplar sets)

For each feature subset, matcher performance is closely correlated with the Good/Bad/Ugly quality assessment, as would be expected from the relationship between quality assessment and minutiae. Most matchers achieve 90-100% accuracy for the Excellent latents, and all matchers show clear decreases in accuracy with respect to quality. Results for the Baseline set were approximately the same as for the Baseline-QA set.

Nearly all matchers improved their Evaluation #1 performance (see Appendix A for the Evaluation #1 results on this set) for nearly every category of quality classification. Notably, most matchers showed greatest improvements in the Bad and Ugly categories over Evaluation #1.

14 Hit / Miss / Loss and Gain Analysis

14.1 Miss Analysis

Figure 12 compares the ability of the matchers to correctly identify latents with the latent examiners' determinations. The figure shows examiner determinations of No Value / Unusable or value for exclusion only (Limited Value), combined with inconclusive determinations from the subsequent examiner review, indicating cases in which either the initial or reviewing examiner determined individualizations were not appropriate.

17.8% of the latents in the test were missed by all matchers at rank 1, 56.8% of which could not be individualized by a certified latent examiner. The initial or reviewing examiners determined that 17.6% of the latents in the test were determined by examiners to be of No Value / Unusable, of value for exclusion only ("Limited"), or resulted in an inconclusive determination; 42.6% of these could be matched by one or more matchers at rank 1.

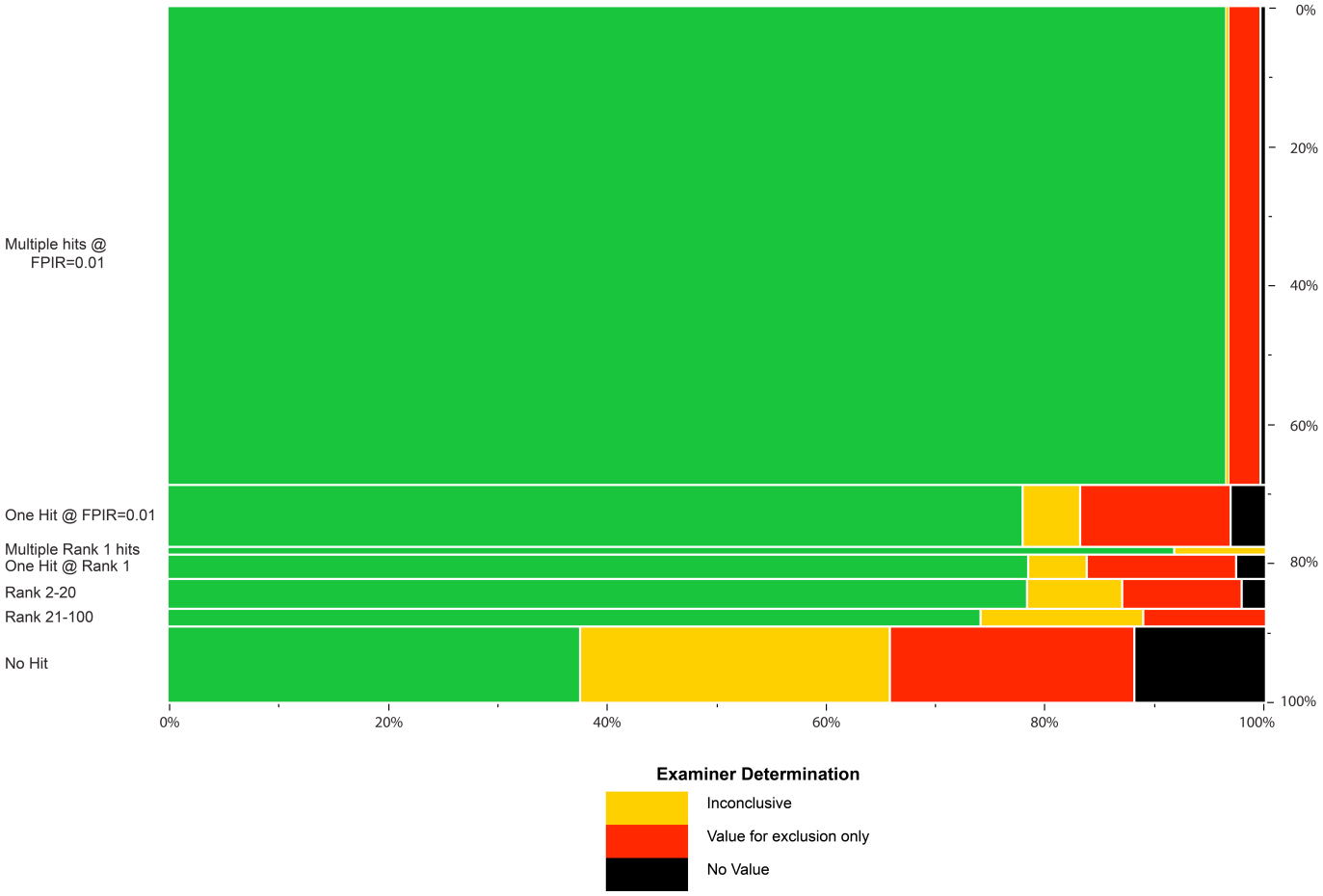

Figure 12: Comparison of matcher hit rates and examiner inconclusive/value determinations, as proportions of Baseline dataset, across all latent subsets.[l]

(Baseline dataset with 10 latents without mates excluded, 1,066 latents — 100,000 rolled+plain 10-finger exemplar sets)

Latents that are No Value / Unusable or that result in an inconclusive determination cannot be used by a latent examiner to make an individualization, and therefore identification of such latents by the matchers cannot be expected to result in individualization decisions. However, it is possible that these algorithmic identifications could yet be of value as an investigative tool.

Table 20 shows the proportion of each latent subset that was missed by all matchers at rank 1. For both Baseline and Baseline-QA, the LG subset had by far the greatest proportion of missed-by-all latents; the other subsets were similar to each other, with LD/LE having the lowest proportion. Table 20 also shows the average minutiae count of the missed-by-all cases: these are relatively consistent, and contrast to the overall average minutiae count for Baseline of 22.5.

[l] *Each latent search (y-axis) is categorized based on the number of matcher hits, rank and score. The highest applicable category is used for each latent search: e.g., a latent that hit multiple times (across all subsets and matchers) at a score equivalent to FPIR 0.01 is included in the "Multiple hits @ FPIR=0.01" category, and the subsequent categories are only for those cases that did not results in multiple such high-scoring hits. Inconclusive cases of no/limited value are indicated by the value determination.*

Table 20: Proportion of data missed by all matchers at rank 1, with average minutiae
(100,000 rolled+plain 10-finger exemplar sets)

	% of data missed by all	Avg. minutiae
Baseline		
LA	23.4	12.2
LE	19.6	11.9
LG	43.3	13.1
All latent subsets	17.8	11.5
Baseline-QA		
LA	25.6	12.3
LB	25.1	12.5
LC	25.1	12.2
LD	21.5	11.4
LE	21.8	11.6
LF	24.4	11.9
LG	45.9	13.2
All latent subsets	16.5	11.3

The distribution of orientation for the missed-by-all cases was not notably different from the overall distribution.

As expected, the latents missed by all matchers include a greater proportion of low-quality prints (Ugly, Limited, or No Value / Unusable) than the overall distribution, as shown in Figure 13. However, the latents missed by all are not solely limited to these cases, and do include Good prints (though none assessed as Excellent). It should be noted that the misses may be due to a lack of good overlap with the exemplar and/or poor exemplar quality in the overlap area, as indicated by the "inconclusive" category in Figure 12.

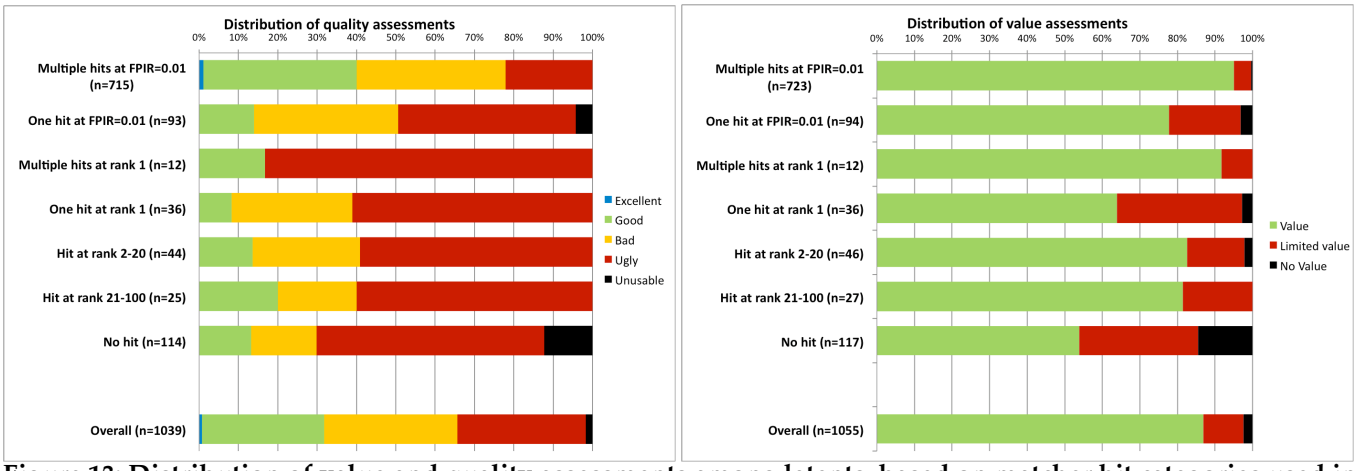

Figure 13: Distribution of value and quality assessments among latents, based on matcher hit categories used in Figure 12
(100,000 rolled+plain 10-finger exemplar sets)

14.2 Hit Analysis

Table 21 and Table 22 show the collective rank-1 "hits" (identifications) made by any matcher (i.e. the union of all latents hit) broken down by minutiae count with respect to examiner-assessed value and quality. For all latents in the complete Baseline dataset, as well as for each quality category, the collective rank-1 hit rates are broken down into search subsets LA, LE, and LG, as well as "any search subset" (i.e. the union of all search subset hits). Adjacent cells representing the search subsets LA and LE are highlighted in green for cases where LE exceeded LA by more than 5%, and highlighted in yellow for cases where LA exceeded LE.

The lowest percentage of "hits" were for latents with low minutiae count and poor image quality. For latents with more than 10 minutiae, minutiae count was the most important factor determining the identification rate, with image quality being secondary. For all of the latent feature subsets, 100% of the latents rated as "Excellent" (all with more than 45 minutiae) were identified by one or more matchers.

For low minutiae counts image quality was a significant factor. Examiner markup had no effect or degraded performance for very poor quality (No Value / Unusable) and very low minutiae count latents (1-5 minutiae). In these cases image-only searches (LA) produced the same or better results than examiner markup.

The greatest gains from adding examiner markup to the search (LD or LE), versus the image alone (LA), were for poor quality (Limited Value and Ugly) latents having 1-30 minutiae.

Table 21: Rank-1 identification rates by any matcher, by latent value and minutiae count

(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

		All Latents				Value				Limited Value				No Value			
# Min	% data	any subset	LA	LE	LG	any subset	LA	LE	LG	any subset	LA	LE	LG	any subset	LA	LE	LG
1-5	5.5	33.9	22.0	32.2	0.0	33.3	0.0	33.3	0.0	40.0	25.7	40.0	0.0	23.8	19.0	19.0	0.0
6-10	15.4	62.8	54.9	59.1	15.9	64.7	58.8	62.7	20.6	57.9	45.6	50.9	7.0	50.0	50.0	50.0	0.0
11-15	19.6	78.9	69.4	76.6	41.6	79.2	68.8	77.1	43.2	76.9	76.9	76.9	23.1	50.0	50.0	50.0	50.0
16-20	15.7	83.8	79.0	83.2	55.1	83.5	79.7	82.9	55.7	87.5	62.5	87.5	37.5	-	-	-	-
21-25	12.7	93.3	88.9	90.4	70.4	93.2	88.7	90.2	70.7	-	-	-	-	-	-	-	-
26-30	9.5	93.1	91.1	92.1	82.2	92.9	90.8	91.8	81.6	-	-	-	-	-	-	-	-
31-35	5.5	100.0	98.3	100.0	91.5	100.0	98.3	100.0	91.5	-	-	-	-	-	-	-	-
36-40	4.3	93.5	93.5	93.5	91.3	93.5	93.5	93.5	91.3	-	-	-	-	-	-	-	-
41-45	3.4	100.0	94.4	97.2	100.0	100.0	94.4	97.2	100.0	-	-	-	-	-	-	-	-
46-106	8.4	100.0	100.0	100.0	98.9	100.0	100.0	100.0	98.9	-	-	-	-	-	-	-	-

Table 22: Rank-1 identification rates by any matcher, by latent quality and minutiae count

(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

	Excellent				Good				Bad				Ugly				Unusable			
# Min	any subset	LA	LE	LG	any subset	LA	LE	LG	any subset	LA	LE	LG	any subset	LA	LE	LG	any subset	LA	LE	LG
1-5	-	-	-	-	50.0	50.0	50.0	0.0	50.0	33.3	50.0	0.0	36.1	22.2	36.1	0.0	15.4	15.4	7.7	0.0
6-10	-	-	-	-	60.0	53.3	60.0	33.3	67.4	63.0	65.2	17.4	62.5	52.1	58.3	13.5	50.0	50.0	25.0	0.0
11-15	-	-	-	-	92.6	88.9	88.9	51.9	85.5	75.4	82.6	55.1	72.9	61.7	71.0	31.8	0.0	-	0.0	0.0
16-20	-	-	-	-	76.5	76.5	76.5	61.8	94.4	90.3	93.1	63.9	74.6	66.1	74.6	39.0	-	-	-	-
21-25	-	-	-	-	92.9	85.7	92.9	81.0	95.2	93.7	92.1	68.3	91.3	82.6	82.6	52.2	-	-	-	-
26-30	-	-	-	-	92.0	92.0	92.0	86.0	97.1	94.1	94.1	79.4	86.7	80.0	86.7	73.3	-	-	-	-
31-35	-	-	-	-	100.0	100.0	100.0	100.0	100.0	100.0	100.0	88.0	100.0	100.0	100.0	50.0	-	-	-	-
36-40	-	-	-	-	92.9	92.9	92.9	89.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	-	-	-	-
41-45	-	-	-	-	100.0	96.4	96.4	100.0	100.0	87.5	100.0	100.0	-	-	-	-	-	-	-	-
46-106	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.6	100.0	100.0	100.0	100.0	-	-	-	-	-	-	-	-

15 Loss/Gain Analysis

When comparing searches based on different latent subsets, a net improvement in accuracy does not necessarily mean that each separate search increased the likelihood of an identification. Figure 14 shows, for successive search runs of the same latents: the number of latents ***not*** identified at rank-1 ("losses"); the number of latents correctly identified at rank-1 ("gains"), and the sum of the two (i.e. "net change" in the number of rank-1 identifications) when additional information (image or feature data) is used for one of the search runs being compared.

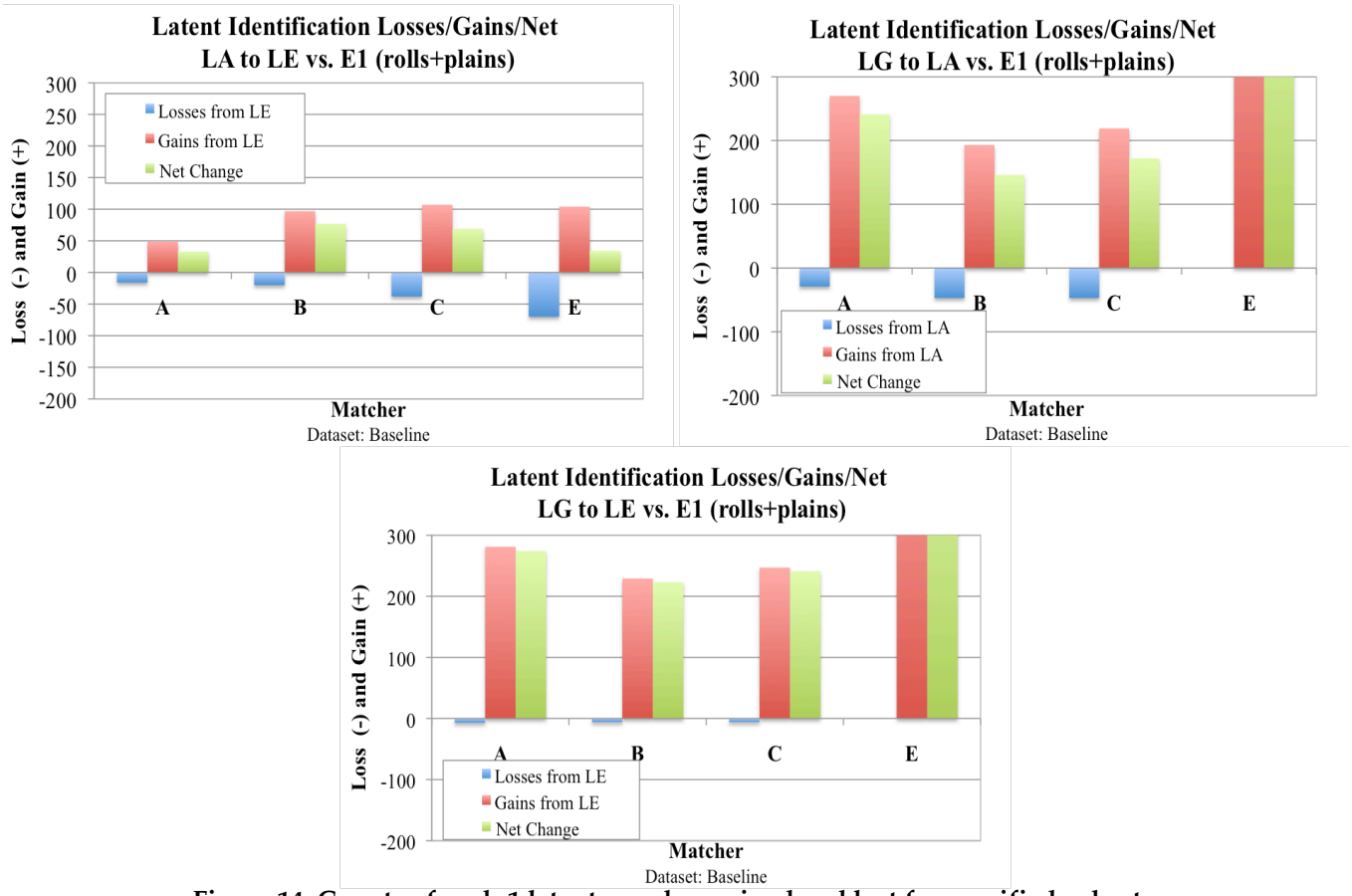

Figure 14: Counts of rank-1 latent searches gained and lost for specified subsets
(Baseline dataset, 1066 latents — 100,000 rolled+plain 10-finger exemplar sets)

In most cases additional image and/or feature data resulted in a net gain in the number of latents identified which is reflected in the identification rate increases discussed in previous sections. However, it is notable that in almost every case there are some latents which were ***only*** identified by supplying less data, rather than more, to the matcher. This unexpected behavior for a portion of the data warrants further investigation.

Little to no commonality exists amongst the losses for differing feature subsets between matchers.

The participants may consider improvements to their algorithms that address the variations in performance levels due to different encoding possibilities.

Even though LA typically far outperforms LG, there are still a number of latents which are matched with LG but not with LA. LG has the least total information, however, when compared with other datasets still produces some hits that are lost with additional information.

If this behavior persists, it may be worthwhile for latents of high importance to be submitted as multiple combinations of input data, e.g. an image-only and image plus feature set searches.

16 Results and Conclusions

1. The highest accuracy for all participants was observed for searches that included examiner-marked features in addition to the latent images.
2. Image-only searches were more accurate than feature-only searches for all matchers.
3. The top performing matchers showed a strong ability to filter out a substantial proportion of false candidates by match score, trading off a moderate drop in accuracy for a very substantial reduction in examiner effort. Since score-based results are more scalable than rank-based results, they provide a better indication of how accuracy would be affected by an increase in database size. This capability could provide important operational benefits such as reduced or variable size candidate lists and greater accuracy for reverse latent searches (searches of databases containing unsolved latents) where a score threshold is used to limit candidate list size.
4. In almost all cases, additional features resulted in accuracy improvement, highlighted in green; cases in which additional features resulted in a decline in accuracy, highlighted in yellow, may be indicative of implementation issues.
5. The ground truth (GT) markup method, in which all exemplar mate images were consulted when marking latent features, yielded an increase in performance over the original examiner markup of about 4 to 6 percentage points for image + full EFS searches, and about 12 to 15 percentage points for minutiae-only searches. Though this method is obviously not practical operationally, it shows that matcher accuracy is highly affected by the precision of latent examiner markup, especially in the absence of image data.
6. Latent orientation (angle) has an impact on matcher accuracy. When the orientation of latents was unable to be determined by an examiner, the rank-1 identification rates were substantially (on the order of 20 percentage points) lower than for the latents for which orientation could be determined.
7. Matcher accuracy is very clearly related to the examiners' latent print value determinations, with much greater accuracy for latents determined a priori to be of Value. The matching algorithms demonstrated an unexpected ability to identify low feature content latents: Matcher A's rank-1 accuracy for No Value latents was 20% on image-only searches, and 34.5% on Limited Value latents.
8. The performance of all matchers decreased consistently as lower quality latents were searched, with respect to the subjective scale of "Excellent", "Good", "Bad", "Ugly", or "Unusable".
9. Analysis showed that the greatest percentage of the misses were for latents with low minutiae count, and those assessed by examiners as poor quality ("Ugly"), "No Value" or "Unusable." Algorithm accuracy for all participants was highly correlated to the number of minutiae.
10. At rank 1, 17.8% of the latents in the test were missed by all matchers. Nearly half of these could be individualized by a certified latent examiner.
11. The initial or reviewing examiners determined that 17.6% of the latents in the test could not be used for individualization (Unusable, No Value, of value for exclusion only, or resulted in an inconclusive determination). Nearly half of these were matched by one or more matchers at rank 1.
12. The highest measured accuracy achieved by any individual matcher at rank 1 on any latent feature subset (excluding the use of ground truth markup) was 71.4%, even though approximately 82% of the latents in the test were matched by one or more matchers at rank 1. This indicates a potential for additional accuracy improvement through improved algorithms. The differences in which latents were identified by the various matchers also points to a potential accuracy improvement by using algorithm fusion.
13. All matchers lost or gained a small number of hits as a function of the feature subset used. For high priority cases, where maximum accuracy is desired, it may be worthwhile to submit the search as separate searches using different levels of EFS, (e.g. search using image-only and again search using image+features, and fuse the results).
14. All matchers were more accurate using galleries of both rolled and plain impressions compared to galleries of either rolled or plain impressions separately. And the use of plain impressions in the gallery compared to rolled impressions resulted in a drop in accuracy. For example, searches that included examiner-marked features in addition to the latent images were approximately 6 percentage points more accurate using combined rolled and plain impressions than rolled impressions alone. Similar searches of plain impressions

alone were (excluding one outlier) approximately 8 percentage points less accurate than searches of rolled impressions alone.

15. The proportion of the total identifications made by a given matcher that were recorded at rank 1 (IR rank-1 / IR rank-100) is an indication of scalability of performance because identifications at higher ranks are less likely as gallery size increases. For matchers A, B and C, 87-90% of total identifications are recorded at rank 1 for subset LA; 89-92% of total identifications are recorded at rank 1 for subset LE; 74-79% of total identifications are recorded at rank 1 for subset LG. These results indicate a potential for developing viable candidate list reduction techniques.
16. The "AFIS" markup approach, in which debatable minutiae were removed, was counterproductive in all cases. This result indicates that the matchers are relatively robust when processing debatable minutiae.

References

1 M Indovina, RA Hicklin, GI Kiebuzinski; *ELFT-EFS Evaluation #1 - An Evaluation of Automated Latent Fingerprint Identification Technologies: Extended Feature Sets; NISTIR 7775*; March 2011. (http://biometrics.nist.gov/cs_links/latent/elft-efs/NISTIR_7775.pdf)

2 American National Standard for Information Systems; *Data Format for the Interchange of Fingerprint, Facial, & Other Biometric Information*; NIST Special Publication 500-290; ANSI/NIST-ITL 1-2011; November 2011. (http://www.nist.gov/itl/iad/ig/ansi_standard.cfm) *[Note: Relatively minor revisions to the EFS portions of this specification have been made since the draft version used in the test, but these did not affect the fields used in ELFT-EFS.]*

3 Hicklin; "Guidelines for Extended Feature Set Markup of Friction Ridge Images" ; Working Draft Version 0.3, 12 June 2009. (http://fingerprint.nist.gov/standard/cdeffs) *[Note: This document has been formalized in "Markup Instructions for Extended Friction Ridge Features", version 1.0, March 2012 (http://www.noblis.org/interop).]*

4 Federal Bureau of Investigation Criminal Justice Information Services; Electronic Biometric Transmission Specification (EBTS) (https://www.fbibiospecs.org/ebts.html)

5 Indovina, et al; *ELFT Phase II - An Evaluation of Automated Latent Fingerprint Identification Technologies; NISTIR 7577*; April 2009. (http://fingerprint.nist.gov/latent/NISTIR_7577_ELFT_PhaseII.pdf)

6 Wavelet Scalar Quantization (WSQ) Gray-Scale Fingerprint Image Compression Specification (https://www.fbibiospecs.org/docs/WSQ_Gray-scale_Specification_Version_3_1.pdf)